Predicting Eye Fixations with a Deep Reconstruction-Based Approach

Francisco Leitao

Fourth Year Project Report

School of Informatics

University of Edinburgh

2019

Abstract

Our gaze is constantly shifting to locations of interest. Throughout the last decades, a lot of effort has been put into deciphering the mechanisms underlying these shifts. In this work, we will attempt to give a brief overview of such research and we will propose an unsupervised solution for saliency modeling that is inspired by the idea of predictive coding and that takes advantage of the existing solutions in the realm of deep learning.

We will describe the implementation of this model and present the iterations that were carried out in order to perfect it. In the end, we will provide qualitative and quantitative results supporting the efficiency of our model and we will highlight some ideas for further improvements.

Acknowledgements

Many thanks to my supervisor, Richard Shillcock, and to his student, Beren Millidge, who were always willing to give a helping hand.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Francisco Leitao*)

Contents

Chapter 1

Introduction

The question of perception is of paramount importance when it comes to understanding how humans function. It has been a topic of philosophical enquiry since time immemorial, but despite all the efforts to unearth its mysteries, we're still very far from having a complete picture of how we process the external world. As tempting as it could be, this work will not focus on the philosophical implications of characterizing perception, as this would lead us to a tortuous path of endless theories and open questions, which is far from the goal of an Informatics dissertation project.

We're going to direct our efforts instead to the problem of understanding the computational mechanisms underlying visual processing.

1.0.1 Structure

In order to do this, we're first going to conduct a literature review of important concepts and insights related to this area, which is the crux of chapter 2. Next, we aim to create a model that integrates that information and attempts to predict eye fixations from raw images. This will be detailed in chapter 3, where we will expand on the different design decisions that were carried out and present the results that were obtained. In the last chapter, we will discuss these results, highlight some limitations about our model and think about possible avenues for further improvements and future work.

1.0.2 Motivation

Visual processing is a fascinating field of study with a surge of recent advances that is propelling the field into unexplored realms. Understanding the nature of visual saliency is important in its own right, as it seeks to uncover the inner-workings of our most refined sense. It is also central to a wide spectrum of other tasks such as computer vision, robotics, image processing etc.

With the advent of deep learning models and the growth of reliable large-scale eye fixation datasets, a lot of solutions nowadays are based on learning the patterns of the provided fixation data and adding a collection of *ad hoc* components to maximize accuracy. While this might be very useful for a range of different applications, it disregards the neuroscientific side, in which one tries to emulate the known functional aspects of the brain.

This work is not an attempt to improve the accuracy of deep supervised models. It is also not trying to create a model that is constrained by the physical properties of the brain. It lies somewhere in the middle, where the final goal is to be inspired by the neuroscientific literature while leveraging the representational power of deep models, in order to create a model that accurately predicts eye fixations.

Chapter 2

Background

2.1 Biological Overview

Vision is the most complicated and delicate sensory modality. It is estimated that more than one fourth of the human brain is dedicated to visual processing [29]. Converting light energy into a clear image that is assimilated in our conscious perception is a very complex task, involving many different stages of processing (Figure 2.6).

Vision begins with light entering the eye through the cornea and being projected onto the retina. Here, around 125 million cells convert the energy into a neuronal signal. They are the photoreceptors and are divided into rods and cones. Rods represent 95% of the photoreceptors and are better adapted to deal with dim light, while cones are meant to transduce bright light and can pick up more fine-grain details for tasks that require a lot of visual acuity. Vision is much sharper in the center of the retina, as it contains many more cones than in the periphery of the retina. The fovea is the small central point where cones are the most densely packed. These photoreceptors are connected to interneurons that relay the original signals to the ganglion cells by activating action potentials. Each ganglion cell receives input from a different number of photoreceptors, from just one or a few in the central region to a larger number in the peripheral areas, which also explains why our peripheral vision has a lower resolution. This number relates to the receptive field of a particular ganglion cell, namely the region of visual space in which the action of light can modify its firing rate. This notion is an important one and we're going to revisit it in later sections. The receptive field of a ganglion cell is divided into a central disk (the centre) and a concentric ring

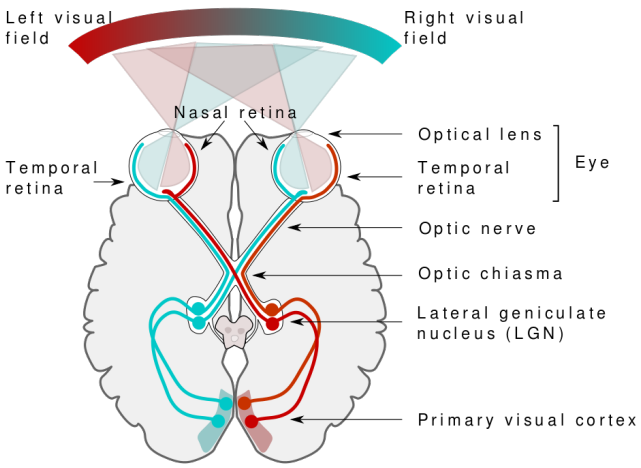


Figure 2.1: The Visual Pathway, taken from [29]

(the surround) that respond in opposite ways to light. For example, if a ganglion cell is activated when light hits the centre of its receptive field, it decreases if light also hits the surrounding area. This center-surround mechanism allows our visual system to maximize the perception of contrast, which is essential for delineating the visual scene and detecting the presence of objects.

The axons of the ganglion cells form a bundle and exit the retina via the optic nerve. Both eyes' optic nerves cross in the optic chiasm, where the nerve bundles from each retina cross paths, resulting in the right half of each eye's visual field being represented in the left hemisphere and vice versa. The signals are then processed in the Lateral Geniculate Nucleus (LGN), where receptive fields are similar to those of ganglion cells, before finally entering the primary visual cortex (V1), or the striate cortex. This first area of the brain is located in the occipital lobe (the back of the brain). Similar to the retina, this region is organized into different layers, with cells possessing more complex receptive fields that identify stimuli like bars and edges of particular orientations [46]. The signal is then transmitted to two distinct pathways, called the ventral stream, which is directed towards the temporal lobe, and the dorsal stream, which is directed to the parietal lobe. The former is known as the "What Pathway" and seeks to identify different forms and objects, whereas the latter is referred to as the "Where Pathway", associated with motion, object locations and connected to motor actions related to the visual field [38] [115]. As the visual information from the primary cortex progresses to these higher levels, receptive fields become increasingly complex. This idea is further explored in the next section.

This simple overview provides enough information for our endeavors. For more information about the anatomical aspects of the visual brain, we refer the reader to [29].

2.2 Representation learning

As we mentioned previously, it is estimated that the retina contains at least around four million cones [88] [82], it has $2^{4000000}$ different configurations (as each can be in either firing or not-firing state). If our brain was to process this with no assumption of redundancy, the amount of information to be considered for later perception would be of the same order of magnitude as the aforementioned monumental figure. Instead, it is much more plausible that our brain takes advantage of the statistical redundancies of the natural world. The visual scenes we encounter on a daily basis are highly interdependent in both space and time, given that the world as we know it is lawful. In this sense, it is hypothesized that the sensory cortical regions of the brain are actually constantly making statistical inferences of external stimuli.

In a famous paper from 1954 [3], the psychologist Fred Attneave asks the question: how would an observer respond to a situation in which the retinal receptors were stimulated quite independently of one another?. To answer this, he attempts to create an image in which every small region is completely independent and uncorrelated from other regions. He presents the following figure, constructed by simply dividing it into 19,600 small cells and filling each cell randomly with either black or white:

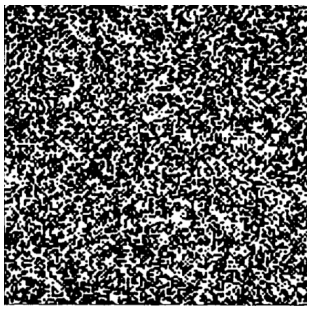


Figure 2.2: taken from [3]

What is striking about this figure is what he terms the “subjective impression of homogeneity” that it gives: it seems that the left half is, in general, very similar to the right half. Since homogeneity relates with redundancy, it is therefore a remarkable

observation, since the image was constructed to be completely non-redundant. This apparent homogeneity in perception is evident in our perception as texture. This points to the fact that "when some portion of the visual field contains a quantity of information grossly in excess of the observer's perceptual capacity, [the brain] treats those components of information which do not have redundant representation somewhat as a statistician treats error variance, averaging out particulars and abstracting certain statistical homogeneities".

In this sense, the goal of the brain is to understand the causes that underlie the received stimuli. In this particular case, our brain infers that the image was constructed with two invariant factors, namely the probability that any cell is either black or white (here, 0.5) and the size of the cells. This computation comes into conscious experience as perceived texture, allowing us to reduce the seemingly chaotic stimulus into simple causes that present a homogeneous whole. The same can then be generalized to objects in the real world, such as trees, chairs or people. All of these are highly complex structures that are presented in radically different forms. We can nonetheless group them effectively based on their holistic structure, meaning our brain is able to reduce such seemingly disparate stimuli, such as a brightly lit almond tree and a red pine in the night time, as being generated from the same underlying cause: they're in fact both trees!

The natural world is filled with such statistical redundancies. By spotting them, our brain is then able to create features and labels that simplify and give meaningful representations to the original data. This is indeed a marvellous feat but we're left with the question of how all this is achieved. Specifically, how is sensory information processed to support higher level tasks?

In 1948, the mathematician Claude Shannon authored a seminal paper [103] in which he set the foundations to what soon came to be known as the field of Information Theory. It aims at quantifying the transmission of information by formalizing it in a mathematical language. The concepts emerging from his work gave researchers the necessary tools to explore possible ways to extract meaningful features from data and thus speculate about the processing of the brain. One such case is the one of Horace Barlow who, in 1961, proposed a theoretical model of sensory coding in the brain that he termed the efficient coding hypothesis. He claims that the action potentials (or spikes) of neurons in the brain form a neural code that minimizes the number of spikes needed to transmit a given signal. In this sense, he predicted that neurons in the visual

system should be optimized to the environment that surrounds them.

Following this, many researchers became interested in understanding how such encoding strategies can allow for the extraction of meaningful representations from raw observations. On this account, the field of unsupervised learning was born and we should spend some time presenting its main concepts and the insights that emerged from the research.

2.2.1 In machines

In broad terms, unsupervised learning attempts to detect predictable structures and model relations between the variables of incoming data. As opposed to supervised learning, these models rely solely on raw observations and are not provided with additional information about the data. Even with the overwhelming prominence of supervised methods today, one should not neglect the importance of unsupervised learning and its relevance for future research. As some of the most prominent names in the field of machine learning put it:

"We expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object" [67]

As we previously alluded, it is important to understand that images are far from being an arbitrary matrix of pixels. In fact, especially with natural images, one can detect a myriad of patterns and complex correlations between pixels. The fundamental task of unsupervised models is then to extract the causes that lead to these patterns, thus transforming the original raw data into more meaningful and useful representations. It is much easier to discriminate between types of trees, for example, when given variables that correspond to properties like colour of leaves and size of the trunk, instead of the original pixel values. This would imply that the features need to be disentangled from one another. Using the same example, the colour of leaves in a tree is an information that is present across multiple pixels of the original image, while it is encapsulated in a single variable in the new higher-level representation. Additionally, we also want the features to be invariant to changes that have little effect on the conceptual structure of the image. Changing the trees position in the image would completely change the values of the pixels but should not produce a big change in its high-level representation.

Achieving this is far from trivial. To begin with, it is essential to pose the problem in terms of statistics, in order to quantify and formalize it. In this perspective, the issue of disentanglement can be restated as an attempt to minimize the redundancy of the original data, or rather to ensure that the output variables are as statistically independent as possible. The resulting representation should also try to keep most of the information of the original data, meaning that it would be possible to reconstruct the image from its high-level representation.

In order to do this, a common strategy is to actually invert the process by learning a way to reconstruct an image from a given representation. These are called generative models and provide a way of generating an image from underlying causes, often referred as hidden or latent variables **h**. The idea is then to infer values **h** that would have generated the data in question using the learned model of reconstruction, with a different inference or recognition model. This step is often called the recognition step and is ultimately what we want to try to achieve. Formally, for a set of parameters $\mathcal{G}$, we can write these opposite procedures as such:

$$p(\mathbf{u}|\mathcal{G}) = \sum_{\mathbf{h}} p(\mathbf{u}|\mathbf{h},\mathcal{G})p(\mathbf{h}|\mathcal{G}) \tag{2.1}$$

$$p(\mathbf{h}|\mathbf{u},\mathcal{G}) = \frac{p(\mathbf{u}|\mathbf{h},\mathcal{G})p(\mathbf{h}|\mathcal{G})}{p(\mathbf{u}|\mathcal{G})} \tag{2.2}$$

where 2.1 refers to the generative model, 2.2 refers to the recognition model and **u** represents the original stimuli.

2.2.1.1 Sparse Coding

Sparse coding [78] is a good example of a generative model. The idea here is to find a sparse representation $H = [h_1,\ldots,h_K], h_i \in R^n$ of the input dataset $U = [u_1,\ldots,u_K], u_i \in R^d$ in the form of a linear combination of a set (or a dictionary) of basic elements $D = [d_1,\ldots,d_n]$, taken from $\mathbf{D} \in R^{d\times n}$. It is thus an optimization problem in which we minimize:

$$\underset{\mathbf{D}\in C, h_i\in R^n}{\operatorname{argmin}} \sum_{i=1}^{K} \|u_i - \mathbf{D}h_i\|_2^2 + \lambda \|h_i\|_1 \tag{2.3}$$

where $||.||_1$ denotes the L1-norm and λ is a regularization parameter. the first term attempts to minimize the reconstruction error while the second term encourages a sparse representation. This problem is optimizing both the dictionary and the coefficients, where fixing one of them leads to solving the other as a convex problem. Most algorithms are then based on iteratively updating one and then the other. In the end, we get a dictionary of basis functions D such that the coefficients in each h_i are as sparse and statistically independent as possible. The recognition step here is then to find h_i such that $p(h_i|u_i)$ is maximized, allowing to represent an image as a linear combination of the dictionary elements. After applying this model, the resulting dictionaries are similar to the ones in the following figure, which is a set of 200 basis functions from 12x12 patches:

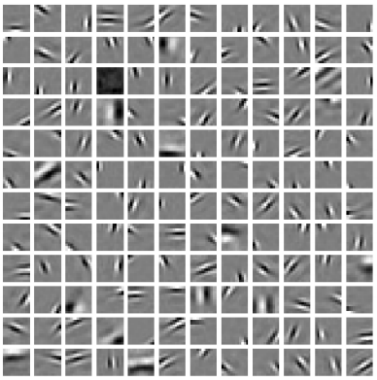


Figure 2.3: taken from [80]

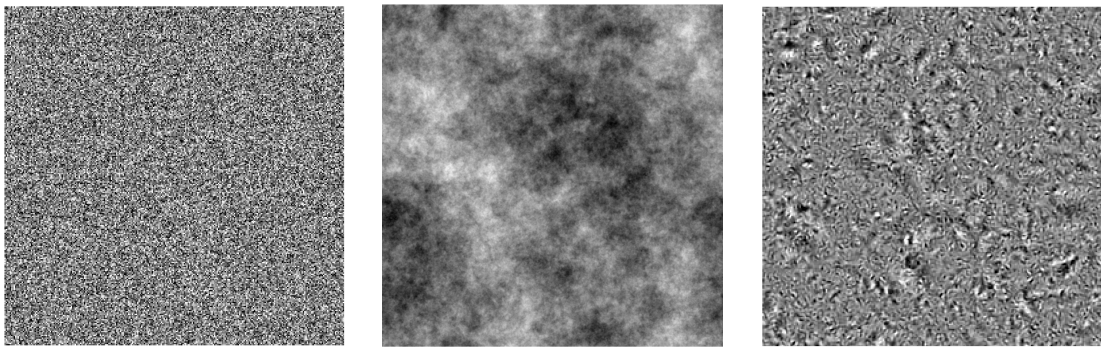
We can see from the above figure that the basis found are similar to Gabor patches and are capable of representing some features of the natural world, such as oriented structures like edges and corners. By describing an image as a combination of such patches, we get a more efficient representation for later stages of processing, since the sparsity allows for outputs that are more statistically independent. Another famous generative model is called Independent Component Analysis (ICA) [8] that tries to maximize the entropy of the output of a linear transformation followed by a non-linearity in order to maximize the mutual information between input and output. The original paper provides a clear explanation of the details of the procedure, but it suffices to say here that it gives very similar results to the ones obtained with sparse coding.

Both these procedures are attempts at finding high-order statistics from the original input, as a means to extrapolate the underlying causes of the image. By identifying such things as edges, it is already inferring the prominence of elongated structures in

the natural world, which requires an understanding of the relationship between pixels that is at a higher-level than first-order (i.e. the mean of intensities) and second-order (i.e. pairwise correlations).

It is known that V1 is highly overcomplete, meaning that there are many more neurons in this area compared to the ones found in the LGN. This implies that neurons in the visual cortex have an exponential firing rate distribution, i.e. at any point in time only a few neurons are firing. In other words, the visual information entering our brain is very sparse. This led researchers to speculate that V1 employs a sparse coding strategy to encode incoming information from the environment [79]. In fact, in [78] they show that learning a sparse code for natural images creates localized, oriented, band-pass receptive fields that are similar to those found in the primary visual cortex.

Nonetheless, a good representation of an image is ideally one in which the causes allow us to synthesize a natural scene. From this perspective, such generative models are still very far from being optimal (2.4). In fact, such a model is supposing that images are generated from low-level causes resembling Gabor patches. Intuitively, we see that a better representation would be one where the causes point to higher-level concepts like objects. The problem here is that these models apply a single transformation to the input, which is certainly not enough to learn all the complex correlations that exist between pixels. Put another way, the outputs of these models are not actually independent, since the structure in natural images comes from underlying causes that are more complex than bars and gratings, affecting the pixels values at higher-orders.



(a) First-order (Mean) (b) Second-order (Variance) (c) Sparse Coding)

Figure 2.4: Image synthesis from low to higher-order statistics, taken from [80]

In order to fully account for the complexities of the relationship between pixels, it is thus more appropriate to think of causes in a hierarchical fashion, in which those at a higher-level of the hierarchy are themselves the causes of causes at a lower level. The abstract concept of a tree, for example, is the cause of both the presence a trunk and some leaves, and these are then the causes of features like textures, contours or colours.

This idea of a hierarchy of abstractions is indeed a strategy employed by the human visual cortex [46] [98]. Hence, if neurons in V1, having small receptive fields, are strongly tuned to rather straight-forward low-level features, such as edges, orientations, sizes and positions, it is not the case with neurons located in higher-levels of the visual pathways. These have increasingly larger receptive fields and are tuned to highly nonlinear properties. For instance, [24] shows neurons in V4 that respond selectively to complex visual patterns such as curved contours and non-Cartesian gratings. Furthermore, neurons in the upper stages of the ventral visual pathway (form occipital lobe to temporal lobe) respond to stimuli such as faces and objects with a high degree of invariance to metrics like position, size or angle. As a caricatured example, [90] verifies this by finding neurons in the medial temporal lobe that only respond when the actress Jennifer Aniston is present in the scene. These highly-specific neurons are generally and informally known as grand-mother neurons. It is therefore important to construct models that are able to encapsulate higher-level information, akin to the neurons present in the higher visual areas.

During the time that these aforementioned methods were developed, there was still no way of training deeper models, as computers still weren't powerful enough for such massive computations. However, [98] introduced a model they called *HMAX*, which implements a hierarchical structure of hand-crafted features, and set the path to the deep learning revolution that was going to happen in the near future.

2.2.1.2 Neural Networks

Most deep models nowadays are part of a bigger family of so called neural networks. Despite its original intention to hypothesize about the brain as is apparent in its name modern neural network implementations are not concerned with trying to emulate real neural mechanisms. Despite this, they are still a very powerful tool and can provide a lot of interesting insights for the field of neuroscience.

A neural network is one of many machine learning algorithms. As is the case with

most of them, there are two main components. The first one is a parameterized model that performs some computation to some input data, while the second one is a learning algorithm that configures the model.

In order to perform the computation, a network consists of layers of units, in which each unit computes a weighted sum of the activations of the previous layers units. In order to obtain a non-linear relationship between the input and the output, each unit applies a non-linear activation function to its output. It is then possible to stack these layers, where the first layer corresponds to the original data while the last layer gives the output. We can thus formulate the computation performed by the intermediate (or hidden) layers as:

$$h_i^l = f\left(\sum_{k=0}^{n} w_{k,i}^l \bar{h}_k^{l-1}\right) \tag{2.4}$$

where h_i^l is the i^{th} unit value in the l^{th} layer of the network, $w_{k,i}^l$ is the weight of the k^{th} unit value in $(l1)^{th}$ layer in the calculation of the weighted sum for the i^{th} unit value in the l^{th} layer, n is the number of units in the $(l-1)^{th}$ layer and $f()$ is the non-linear activation function. This last step is an important one, since without it, stacking layers would still lead to a simple linear transformation of the input.

In order to learn, there is a predefined cost function $C(\theta, U)$, that evaluates the performance of the network given the parameters θ on the data U. It then becomes an optimization problem in which the goal is to minimize said function by iteratively computing the gradient of the cost function on a batch of U, and then using stochastic gradient descent (SGD) to navigate the functions landscape and update the parameters by taking small steps η in the direction of the gradient:

$$\begin{aligned} \Delta\theta &= \eta \frac{\partial C(\theta, U)}{\partial \theta} \\ \theta &\leftarrow \theta - \Delta\theta \end{aligned} \tag{2.5}$$

Note that variants of this procedure are normally used in practice, where η is dynamically adjusted for better convergence [124] [57]. For networks with multiple layers, the gradient needs to be computed using a technique called back-propagation [39].

2.2.1.3 Auto-encoders

The cost function is easy to formulate for supervised learning problems, by trying to have the output match the provided class label for a given input. It is not as straightforward in the case of unsupervised learning. In most cases, the objective is set by a cost function that corresponds to the reconstruction error of the generative model, thereby making sure that the network can generate the original data from the inferred causes of the recognition model. This is also the case with sparse coding, where the first term of equation 2.3 computes the Euclidean distance between the original image and its associated representation as a linear combination of the dictionary elements. This idea motivated the development of auto-encoders, a neural network in which a generative model is appended to a recognition model that computes the latent variables **h** that are then used by the generative model to synthesize the original image **u**:

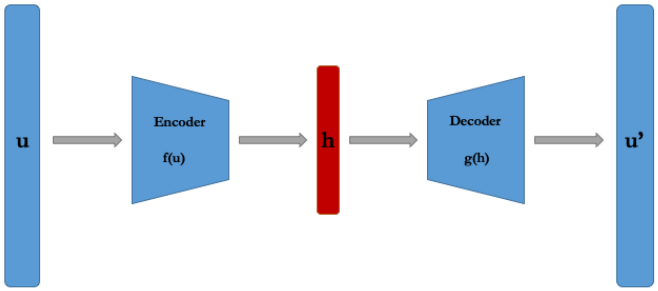


Figure 2.5: Diagram of an auto-encoder

Given this model is a deterministic one, meaning it doesn't operate within a probabilistic framework, we use the terms encoder and decoder as opposed to generative and recognition model.

The goal is for both functions to be each others inverses, i.e. $g(f(\mathbf{u})) = \mathbf{u}$. In order to accomplish this, the previously mentioned strategy is employed, in which a cost function measures the Euclidean distance between the original image and the reconstruction:

$$C_{error} = \sum_{\mathbf{u} \in U} \|\mathbf{u} - g(f(\mathbf{u}))\|^2 \tag{2.6}$$

Both the encoder and the decoder can be composed of a stack of layers, allowing the network to learn non-linear representations. It is, however, important to note that some constraints need to be taken into consideration before the model can learn meaningful representations. Since the measure of performance is solely based on reconstruction accuracy, it is easy for the network to just learn identity mappings, where $f(\mathbf{u}) = \mathbf{u}$ and $g(\mathbf{h}) = \mathbf{h}$. In order to prevent this, a common solution is to have the dimensionality of $\mathbf{h}$ smaller than that of $\mathbf{u}$, forcing the network to compress and decompress the image, and hopefully removing redundancies.

This solution can be sub-optimal, since the creation of a bottleneck doesn't entail that the network will learn semantically interesting features, and forces a dense representation, as opposed to a sparse one. To account for such a problem, a range of possible modifications was put forth. A common one is to remove the bottleneck but add a sparsity constraint to the cost function (like in equation 2.3), encouraging the network to represent a certain image with as few latent variables as possible. Another more recent solution consists of withholding or corrupting parts of the input during training [118] [86] [127]. In [118], they add noise to the input, while in [86], they remove a block of the image, making the network try to inpaint the missing region. Instead of spatial modifications, [77] removes data in the channel direction instead, meaning the auto-encoder takes as input the gray-scale version of the original image and attempts to colorize it. All these solutions dont allow for the possibility of the network to simply learn the identity mapping since the input and the target reconstruction are different. Additionally, by removing large chunks of information, theyre forcing the networks to reason semantically, in order to fill the gaps. This insight is an important one and well get back to it when constructing our model.

One last thing to note before we move on is that, when dealing with images, convolutional neural networks are used most of the time.

2.2.1.4 Convolutional Neural Networks

With the advent of increased computational power, machine learning has taken over the field of computer vision. In order to handle images, the most successful approach has been to use a variation of neural networks called a convolutional neural network (CNN). Here, the units are not connected to every unit in the subsequent layer. Instead, outputs are only connected to inputs in a certain receptive field. This allows the

network to be spatially aware and invariant to translations, and is more related to the connectivity pattern between neurons in the visual cortex, where each neuron responds to stimuli in a restricted region of the visual field, which is its receptive field. It also significantly reduces the number of free parameters, allowing deeper networks to be trained.

CNNs are now ubiquitous, giving state-of-the-art solutions to most vision related problems. It is unlikely such a network uses computational mechanisms that are similar to the human visual system, but its clear that the representations learned are very similar to the ones computed by the brain. In [122], they show that CNNs are highly predictive of neural responses in V4 and IT cortex, which are the top two layers of the ventral visual hierarchy. Similarly, in [21], they are able to compare the performance in object recognition between such networks and primates, and demonstrate that the representations learned using CNNs rival those inferred by the brain.

2.2.1.5 Supervised Versus Unsupervised Approach

Nonetheless, this representational power is mostly seen in supervised models where labelled data guides the network towards finding good discriminative features. As we previously mentioned, the brain doesn't learn through supervision and the question remains as to whether such high-level features reminiscent of the activity of the "grandmother neurons" can be obtained with auto-encoders. This is exactly the question that the authors of [66] set to solve, the answer being that it is indeed possible. By training a deep auto-encoder on an enormous amount of images, they were able to find neurons that respond to high-level features like faces in an invariant way.

This paper was a pivotal one for the further development of unsupervised learning as it showed for the first time that learning from raw data can be enough to get a good representation of the outside world. This said, even with the advent of increasingly better solutions, we still have a long way to go before reaching representation models that match the robustness of our visual system. The question that needs to be asked then is how the brain is able to do it. This largely remains a mystery but a range of different hypothesis emerged recently. We shall spend the next section on giving a brief description of some of these ideas.

2.2.2 In Our Brains

During the 20^{th} century, it was believed that the brain extracted knowledge from sensations, acting as a simple feed-forward mechanism. Today, we know that its doing much more than acting as a "cognitive couch potato", as the philosopher Andy Clark would put it.

Researchers have found a huge number of feedback connections in the brain, action potentials that go from higher-levels of the visual cortex to lower-levels. In this sense, each pair of regions is reciprocally connected, leading to a lot of speculation as to why that is the case.

In 1999, a famous paper by Rao and Ballard [93] set out to explain the phenomenon of the extra-classical receptive field effect of end-stopping. This relates to empirical evidence that shows how some neurons in V1, V2 and V4 can have their activity modulated by stimuli outside their classical receptive field (CRF), responding optimally to a line in a certain orientation but having their firing rate inhibited whenever that line extends beyond the neurons CRF. This was usually justified by the existence of horizontal intracortical connectivity, but the paper shows that this can also explained by claiming that the existing feedback connections in our brain actually try to predict the activity in lower-level areas, while the forward flow of signals convey the prediction errors. They create a simple computational model of this idea that is consistent with neuroanatomy and are able to show that the aforementioned extra-classical effects emerge as a consequence of such an architecture. This idea goes in accordance with the theory proposed in 1992 by David Mumford [75], a principle known as predictive coding. Although this idea has been gaining traction in recent times, its origin can be traced back to the German polymath Hermann von Helmholtz who, in 1860, speculated on the concept of unconscious inference, describing our perception as being largely influenced by our expectations.

Such a model of the brain provides an explanation of what we discussed in the previous sections, namely, how the brain adapts to the statistical properties of the external world. The idea is that only the unexpected output at one stage is transmitted to the next layer. The predictions coming from the feedback connections suppress the neural activity that goes in accordance with the brain's expectations. The prediction error that flows upward is then incorporated by the higher-levels in order to perfect the predictions for the future.

This provides a framework in which we can recognize similarities between the unsupervised algorithms we presented above and the functioning of the brain: the feedforward activity akin to a recognition model and the feedback responses related to a generative model. What's more, Rao and Ballard's framework essentially unifies both the task of inference and that of learning under one underlying paradigm [33].

It also hypothesizes that perception is largely a generative procedure, rather than a mere transformation of incoming stimuli. This would explain many phenomena other than the extra-classical effects presented in [93]. Dreams, for example, could be a result of such a generative procedure in the face of no incoming data. It could also explain why were constantly performing microsaccades, which are short and involuntary movements of our eyes. Since images that are kept fixed in the same point of the retina rapidly fade away [35], these movements would reintroduce error and thus reactivate our visual cortex activity to process the input [74]. Furthermore, there is the phenomenon of Mismatch-Negativity (MMN), wherein an odd stimulus presented in an otherwise repetitive sequence of stimuli triggers an event-related potential (ERP), a strong neuronal response measured by means of electroencephalography (EEG), even in the absence of conscious attention. This response can indeed be seen as a prediction error signal [108] [109]. This effort in explaining some of our behavioural and neurophysiological idiosyncrasies under the light of predictive coding led some to speculate on whether such a view could be the cause of our internal emotional states [7] and some even went as far as to claim that it could provide a neuroscientific foundation for the controversial Freudian formulations of our psyche [20].

Under this idea that there is a generative process in our brain that provides an expectation about the outside world, we can describe perception as a Bayesian inference model. In general terms, given a hypothesis H and some incoming data D, our brain would update the probability $P(H|D)$ following Bayes' Rule:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{2.7}$$

In the visual case where a free-viewing scenario is considered, H corresponds to features of a stimulus (e.g. size of object, orientation of edges), while D represents the incoming raw visual input. Let's think about what each term from the above formula represents:

- In the Bayesian vernacular, $P(H)$ is the prior distribution. In this case, it corresponds to an individual's prior expectation about the probability of encountering the features expressed by the hypothesis H. This distribution should be optimized to the statistical properties of the visual world experienced by the individual.
- $P(D|H)$ refers to the likelihood of observing D given H, measuring how compatible the incoming data is with the hypothesis.
- $P(H|D)$ is the final posterior distribution that provides the probability of a certain set of features given the stimuli. When the function that infers the hypothesis from the data is non-linear (which is certainly the case in visual processing), this distribution is computationally intractable. It is assumed then that the brain computes an approximation of this Bayesian inference. In [10], the author presents a simple mathematical tutorial that explores plausible ways for the brain to perform such computations.

This provides a probabilistic framework where the descending connections of the cortical hierarchy represent a generative model defined by the prior information and the likelihood. A comparison between the model's output and the actual sensory input yield a prediction error that updates the generative model to better predict input in the future.

The probabilities allow us to consider input as a noisy source of information, where the goal of the brain is to reduce such noise. When the data becomes more uncertain (given low luminance or new visual experiences), the prior information should have a bigger influence on the interpretation of the data. The ultimate goal of the brain can be to reduce this uncertainty which requires the prior to reflect the statistics of the sensory world as accurately as possible. In [9], they show empirically that this is the case in ferrets. As they grow older, the spontaneous activities of their brains when faced with no visual stimuli gets gradually closer to the activity evoked when viewing natural scenes, implying that they develop an increasingly optimal model of the world as they age. This plasticity is said to be a universal property that applies not only to the visual cortex, but to all cortical areas. In [104], they show that the primary auditory cortex can learn to process visual information like V1, by rewiring the auditory cortex of ferrets with nerves from the eyes.

This idea of reducing uncertainty was formalized and generalized to all living organ-

isms in a unifying paradigm called the free-energy principle [30] [28] [31]. The author of such papers even claims that consciousness could be seen as the constant interplay between our beliefs and perceptions [100] [32], but such radical conclusions are still wildly speculative and face a lot of criticism.

2.3 Modelling Saliency

2.3.1 Visual field and visual attention

From the previous discussion, it seems clear that seeing is much more than just passively receiving external stimuli. Every second, our eyes receive around $10^8 - 10^9$ bits [13], making it impossible to assimilate it all into conscious experience. Given the complexity of visual scenes, our brains must efficiently deal with monumental amounts of stimuli by selecting a subset from all of the available information for further processing.

This view is beautifully emphasized in Jorge Luis Borges' *Funes el Memorioso*. It portrays the life of a man with absolute memory and perception. One could think that this constitutes an advantage, but the author would disagree. In fact, the character becomes unable to think and see. By perceiving everything, he is unable to select and interpret the visual world; it leads to a form of blindness. In more prosaic terms, we can't make inferences over the whole visual field in parallel, since this would require a tremendous amount of computational resources [25]. In order to successfully navigate the world around us, we need a selective attention that allows us to prioritize some information while ignoring the rest; this is what is meant by visual attention. This disregard for most incoming stimuli is empirically demonstrated by change blindness, in which big changes in an observed image are ignored by the viewer in a natural viewing condition, but can be noticed if directed to them [105] [54]. The ubiquity of attention in our everyday lives makes it an immensely interesting field of study and many researchers have been drawn to it in the past decades. A quick PubMed search with the keyword of "visual attention" yields 2,400 articles addressing the subject since 1980, with half of them having been published since 2005. This said, we are still very far from having a consensual understanding of its neural underpinnings and even of a precise high-level description.

If the human visual field is a canvas spanning 135 x 220 degrees, one should note that only 1% of it is registered by the fovea [116]. The narrow, high-resolution foveal visual field represents nonetheless 10% of the information sent to the brain and there is a steep decline in visual acuity from the foveal region to the peripheral regions of the visual field. This makes it so that the foveal region is generally considered to be the focus of our attention. After selecting a region of interest, the information is placed on the fovea through quick movements of the eyes called saccades. The information is then relayed to the brain that further processes this region. But what is the mechanism that selects the place in which we want to turn our attention to?

The answer to this question is separated into two main components: One works in a purely bottom-up stimulus-driven manner, identifying elements in the visual field that stand out, while the other one is a top-down task-dependent mechanism, driven by the intentions of the viewer and his previous experience, working in a slower, more deliberate way [112]. The relative influence of both these components to reach visual consciousness is variable and still not properly understood. It is nevertheless accepted that they work in tandem - with both competition and collaboration - and it has been shown that almost no search is driven purely by bottom-up or top-down mechanisms [27]. It seems obvious that the former takes place in the earlier visual pathways and that the latter has an influence through interactions in higher-levels of the cortex; this has indeed been found to be the case [73]. For this reason, it is much easier to model the bottom-up phenomena, since they correspond to well-known, fairly simple mechanisms in the retina and the LGN which are universal to every human being, whereas top-down influences interact in highly complex ways, specific to the configuration and current state of each individual. Consequently, many researchers have turned to the easier problem of modelling the bottom-up processing of visual stimuli, and we should dedicate the following paragraphs to understand it better.

Before we do, it seems intuitive that we can integrate this within the perspective of predictive coding discussed previously, with bottom-up processes related to feed-forward mechanisms and top-down control associated with the feedback activations that represent the brain's expectations [2]. While this might be true to some extent, the brain's complexity wouldn't allow for such a simple answer. Even the top-down/bottom-up duality seems to give an incomplete account of things [56], and can possibly be a problematic concept altogether [94]. The topic is in fact the subject of heated discussions among neuroscientists [5] [95], but it is important to point out that this dichotomy and

its functioning within a hierarchical process is still a useful conceptual simplification. We refer the interested reader to [110], which provides a good discussion that seeks to understand this relationship between attention and expectation.

2.3.2 Defining Saliency

In the simplest terms, saliency corresponds to the regions that stand out in a scene, be it a person, an object, a pixel or anything else. The task of quantifying this qualitative description is at the heart of the research that has been made about the subject. Given that the human attentional mechanism is affected by such saliency detection, it is a valuable field of study for modelling human behaviour in the interaction with ones visual surroundings. In fact, we constantly shift our focus to salient regions, in a quick and efficient process of selection that stems from an evolutionary history of rapidly detecting possible preys, predators or other dangers in the visual world. Attention can thus be captured by some stimuli, automatically overriding any cues coming from top-down search tasks [6] [71]. A red apple among green leaves or a flickering light in an otherwise static scene attract our visual attention by standing out from its surroundings.

Hence, it is postulated that a pre-attentive saliency map is computed in the early stages of the pipeline, meaning the visual field is processed in its entirety to extract simple local features and identify the ones with highest saliency [48]. In [128] and [7], thy identified neural activities in the primary visual cortex (V1) that create such a bottom-up saliency map.

Furthermore, many other studies have corroborated the idea that attention and saliency are deeply connected by showing a strong correlation between eye movements and low-level salient contrasts under natural viewing conditions [84] [60] [85] [111].

Note here that eye fixations are seen as a proxy for attention shifts under the assumption that there is a close link between both, even if they are distinct. In fact, it has been shown that attention precedes eye movement [72], and that there is such a thing as covert attention, which refers to an attentional allocation without an accompanying eye movement. This being said, for the remainder of this work, were going to treat human fixation data as the ground truth for human attention, as is the case in most benchmarking studies [50] [15] [96] [14].

2.3.3 Formalizing Attention as a Computational Problem

Were trying to explore the computational mechanisms underlying attention. This is a vague and wide-ranging topic, where models can be built for a range of different purposes, whether to further elaborate on behavioural aspects related to psychology [43], to build networks that are neurologically plausible [76] or even to provide insights for fields like computer vision [52] [1], image processing [4] and human-computer interaction [41]. This leads to a fragmented inter-disciplinary literature, where some studies seek to understand the properties of the brain and use models to verify hypothesis, while others use already existing knowledge about the visual system to inspire the construction of models for applications in various realms of science and engineering. Furthermore, models of attention can be dynamic and explore eye movement prediction across time [107], while others assess attention as a static computation. In this study, we will refrain ourselves with the latter category, and focus specifically on models that compute saliency maps. We will also not concern ourselves with examining the models' level of isomorphism to biological visual processing, and we'll simply point to some interesting conceptual ideas about the visual system that might help in guiding the construction of efficient models.

For the construction of our proposed models and to compare it to other existing solutions, we will only consider solutions that deal with predicting the first few seconds of eye movements under free-viewing conditions of either natural or synthesized scenes, as opposed to models that focus on visual search tasks (e.g. finding a specific object in a scene) or interactive tasks (e.g. playing a sport). Following the formalization stated in [13], we can then define the problem as such: Given a set of N images $I = \{I_i\}_{i=1}^{N}$ and K subjects who viewed these images, let $L_i^k = \left\{p_{ij}^k, t_{ij}^k\right\}_{j=1}^{n_i^k}$ be the vector consisting of eye fixations $p_{ij}^k = \left(x_{ij}^k, y_{ij}^k\right)$ and the corresponding time at which they occurred t_{ij}^k for the k^{th} subject over image I_i. Let the number of fixations of this subject over i^{th} image be n_i^k. The goal of an attention model is to find a function f mapping the stimulus to a saliency map, which minimizes the error on eye fixation prediction: $\sum_{k=1}^{K}\sum_{i=1}^{N} m\left(f\left(I_i^k\right), L_i^k\right)$, where $m \in \mathcal{M}$ is a distance measure. In section 3.6.1, we will present some of these measures, that we're going to use later for the evaluation of our model's performance.

2.3.4 Early Solutions

The first effort towards the conceptualization of a pre-attentive saliency map to guide visual attention was proposed in 1980 [114], where they speculate that a number of low-level features like colours, shapes or movements are extracted from the scene in an unconscious and massively parallel manner. A focused attention merges such features under a unified master map of locations that allows the individual to consciously register the presence of objects. This was further explored by [59], who first introduced the idea of a two-dimensional saliency map. It represents the combination of features that is then used to identify a winning location corresponding to the most salient region and becoming the focus of the systems attention ("Winner-Take-All"). This location then becomes inhibited and the system shifts to other salient regions in a sequential way. Such a search process is performed to account for what is called an inhibition of return, where an individual disregards already attended regions in order to find novel information in the visual scene [89].

This model was finally implemented in 2001 [49], sparking a surge of interest in the field. They compute three features - luminance, colour and orientation - from different scales of the image, in order to extract local center-surround (CS) contrasts. These then give a set of feature maps across different scales and channels, which are finally linearly combined and normalized to generate a final saliency map. This idea of local contrast as a reason for saliency was inspired by the center-surround mechanisms introduced in section 2.1.

This work cemented the use of center-surround contrasts as one of the most influential methods to extract low-level saliency. Further extensions of this model add other features such as motion [102] or depth [44].

Problem: While such models work decently well, they tend to give poor results on images with textures and symmetrical structures [121] (Figure 3.10). This is due to the fact that such features require the model to reason within a relatively large neighbourhood. It is then essential to devise models that take into consideration the context of a certain location at a higher-scale in the image. The solutions that follow tackle this problem in a few different ways.

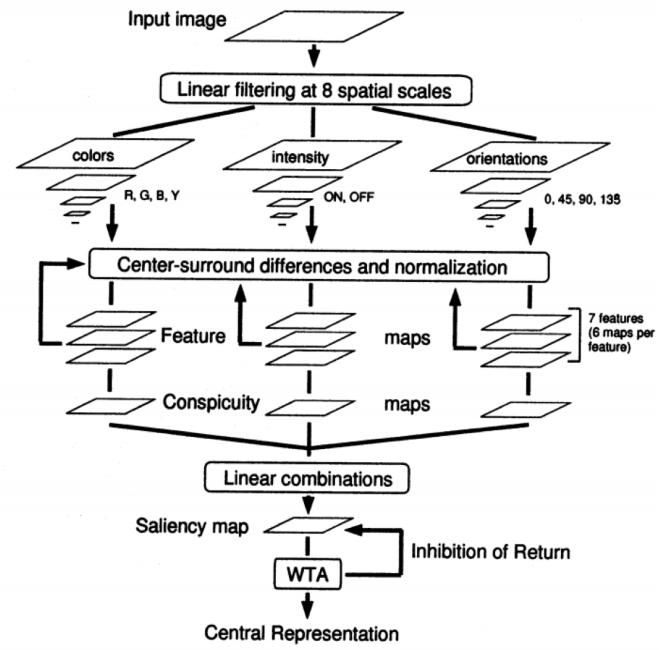


Figure 2.6: Model architecture proposed in [49]

2.3.5 Information Based Models

As we saw before, we can view the brain as a machine that is constantly performing Bayesian inferences. In this sense, models would integrate a prior knowledge (context extracted from spatio-temporal data) and sensory evidence (target features) using Bayes rule, to extract a posterior probability of a certain region in its relationship with the overall scene. In the following figure, given a prior information based on natural scenes and the given image, it's clear that patch A is not surprising at all, as its content can be easily deduced from its surrounding context. To a lesser extent, the same can be said of patch B, given the elongated edge that extends beyond it. Predicting patch C is however very hard, as the surrounding context cannot infer the presence of a tree from the rest of the image.

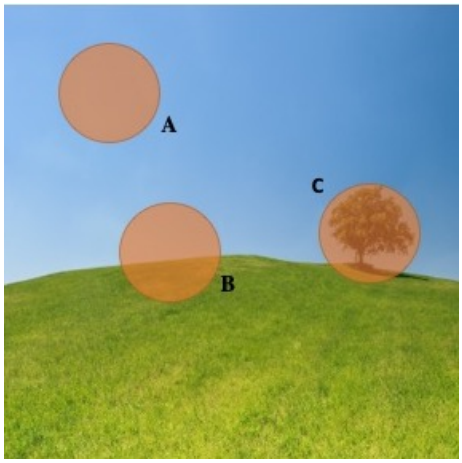


Figure 2.7: Example highlighting how surrounding context is related to surprise

Torralba [113] is the first to introduce this concept and uses it to model visual search tasks. Itti and Baldi [47] explore this idea further and provide a Bayesian framework that attempts to clearly define the notion of surprise. At a qualitative level, a surprising stimulus is one that significantly changes the beliefs of the observer. They quantify it by computing the Kullback-Leibler (KL) divergence between posterior and prior beliefs.

Some researchers extended this idea by constructing probabilistic models that are based on a graphical representation such as Hidden Markov Models (HMM) or Dynamic Bayesian Networks (DBN) [42]. Based on the same idea of surprise, the saliency of a given region can be determined with Shannons self-information measure [16]. These models are inspired by the early insights of Barlow and Attneave presented in section 2.2.1. The goal is to quantify the amount of information of an image patch given the content of its surrounding regions. The self-information of a feature f is $I(f) = -logp(f)$. With this definition, the saliency of a region corresponds to the inverse likelihood of observing f.

However, even for a small image patch $\mathbf{p}$, its probability distribution resides in a high dimensional space that makes the computation of $p(\mathbf{p})$ intractable. In order to overcome this problem, information based models use the techniques presented in section 2.2.1 and represent each patch as a linear combination of basis elements computed with ICA [16] or sparse coding [12], and looking like Figure 2.3. This makes it so that each patch is now represented by a vector $\mathbf{w}$ of N variables (given N basis functions), where each w_i has a value v_i corresponding to the contribution of a particular basis function

in the representation of the patch. The presumed independence of the basis elements allows for a simple computation for the likelihood of a patch $\mathbf{p}$, with:

$$p(\mathbf{p}) \approx p(w_1 = v_1, w_2 = v_2, \ldots, w_n = v_n) = \prod_{i=1}^{h} p(w_i = v_i) \tag{2.8}$$

The estimation of the N-dimensional space is thus reduced to N one-dimensional probability density functions. To obtain these density functions, one then simply considers the distribution of values taken by each w_i in the global context of the surrounding patches.

The interest in such an approach is two-fold. First, applying ICA or sparse coding disentangles the patch representations into more independent features, and second, it allows for a simple way of calculating the dissimilarity between neighbouring regions.

For these methods, the patches used to calculate the current locations rarity differ. In [126], the author computes the self-information of a certain patch as the difference with the entire set of natural image statistics. On the other hand, [113] only uses the current image statistics, as a foreground object is likely to have features that are very different from the background. In [18], the authors place an even tighter constraint and only compute rarity based on the surrounding patches.

It is hard to gauge whether local or global dissimilarities should be used. For the case of Figure 2.7, it seems that local context is enough to give a good measure of saliency. However, it often happens that a local patch is very similar to its surrounding but the whole region is still rare when compared to the full image. By limiting saliency to a local surrounding, it may suppress areas within a homogeneous region that could otherwise be rare. For example, a uniformly textured object would only be considered salient at its borders. In order to overcome this trade-off, [12] propose a unified model that combines both approaches, which had always been treated independently. To do this, they calculate the local saliency S_l of a certain patch $\mathbf{p}_i$ as:

$$S_l^c(\mathbf{p}_i) = \frac{1}{L} \sum_{j=1}^{L} W_{ij}^{-1} D_{ij}^c \tag{2.9}$$

where L is the number of surrounding patches, while W_{ij} is the Euclidean distance between $\mathbf{p}_i$ and a surrounding patch $\mathbf{p}_j$. This ensures that patches that are further away from $\mathbf{p}_i$ have a smaller influence on the computation of its saliency. D_{ij} is the

distance between the sparse coding representations of $\mathbf{p}_i$ and $\mathbf{p}_j$, computed using an L2-distance, for example.

They then compute global saliency S_g using the method of self-information presented in [16]. They finish by combining the results to produce a final saliency map S_{lg}. To do this they simply use:

$$S^c_{lg}(\mathbf{p}_i) = \mathcal{N}(S^c_l(\mathbf{p}_i)) \circ \mathcal{N}\left(S^c_g(\mathbf{p}_i)\right) \tag{2.10}$$

where $\circ$ can be any operation that integrates both values, such as $+$, $-$ or max.

Problem: While these models consider context within a larger-scale of the image, the features they use to compute the dissimilarity with, and thus the saliency, are the ones computed with linear methods like sparse coding. These are still low-level features like orientations and luminance disparities akin to the ones used in [49]. As we mentioned before, causes for attention happen within a richer semantic context. The above models would compute the saliency of a face as its dissimilarity with the surrounding area without ever considering that it's a face. The surge of machine learning solutions helped solve this problem.

2.3.6 Machine Learning Models

Realizing that our gaze tends to focus on high-level concepts like faces, people and text, [51] proposes a model composed by a set of hand-crafted high-level (e.g. faces), mid-level (e.g. horizon lines) and low-level (e.g. orientations) features and trains a linear SVM to assign the weights for each feature. In order to learn these weights, they used a supervised approach in which predictions were compared against ground truth eye fixation data.

With the rise of deep learning models, features no longer need to be hand-crafted, with CNNs being able to learn robust, scale-invariant features without the aid of explicit human intervention. At each layer, a CNN computes a set of feature maps that are combined for the computation of the subsequent layer's output. We can thus think of a CNN that consists of a single convolutional layer and that is followed by a fully connected layer which combines the feature maps, as a generalization of Itti's model [49], where features such as orientation, color or luminance are found naturally by

matching the computation to a given label [125], instead of being hand-picked based on basic cognitive properties of the brain. Stacking such layers allow for deep, non-linear representations with richer semantic understanding. To leverage this representative power, a range of models were proposed.

The first one was eDN (ensemble of deep networks) [117], that combines the output of several shallow convolutional networks and trains these with a linear SVM against true eye fixations. Following this, DeepGaze [63] used a deeper network based on the famous AlexNet architecture [61] and consisting of five layers. A further refinement of the initial model, where low-level and high-level features are combined in a careful way, was later proposed and named Deep Gaze II [65]. Other solutions like SALICON [45], SalGAN [83], DeepFix [62] or ML-Net [23] were proposed. For a thorough listing and explanation of these models, refer to [11]. All seek to leverage the power of CNNs and other new machine learning techniques like Generative Adversarial Networks [91] or Long Short-Term memory modules [36], and then combine them with insights from saliency studies [64] [19], where it was shown that fixations have a strong bias towards the center of the image, for example, or that combining results at different scales of the image can be very helpful [53].

Problem: Despite giving impressive results, these models are still dependent on supervision through eye fixation data. This allows the networks to generate maps that are close to ground truth but it doesn't provide any meaningful insight towards a better understanding of the functional properties of the brain. Additionally, research has shown that current state-of-the-art deep models can still be worse than traditional methods in some cases where low-level saliency is important [92].

Chapter 3

Models

3.1 Conceptualization

Our goal is to conceptualize a solution that takes advantage of the robust representations of deep models while incorporating the conceptual framework of predictive coding. In order to do this, we follow the idea proposed in [121] [120] [68], where rarity of a patch is defined as the reconstruction residual of representing the patch with a linear combination of its surrounding patches, as opposed to using Shannon's self-information.

In the perspective of predictive coding, the reconstruction can be seen as the brain's expectation, with a big error corresponding to a high level of surprise. We aim to create saliency maps, in which the most salient regions are the ones that are harder to reconstruct from a prior expectation incorporated by the model.

In order to do this, we use a deep convolutional auto-encoder, which is just a regular auto-encoder that uses convolutional operations in both the encoding and decoding networks. As we said in section 2.2.1.4, this allows the model to consider the spatial structure of the input and learns features that are spatially invariant. We keep the patch-based representation used in section 2.3.5. We want the network to reason about the relationship between each patch and the rest of the image. In order to accomplish this, we input the original image with the selected patch removed. We then compare the resulting reconstruction with the original patch and calculate the error between them. The higher the error, the higher the saliency in the region of the patch. In order to learn the parameters of the network, the training is performed in a similar fashion, where a

random block is removed from the image and the error is computed as the Euclidean distance between the original block and its reconstruction, like in equation 2.6. The gradient of this error function is computed relative to the parameters of the model using backpropagation and the parameters are then updated to navigate the error landscape towards the opposite direction as that of highest gradient, like in equation 2.5.

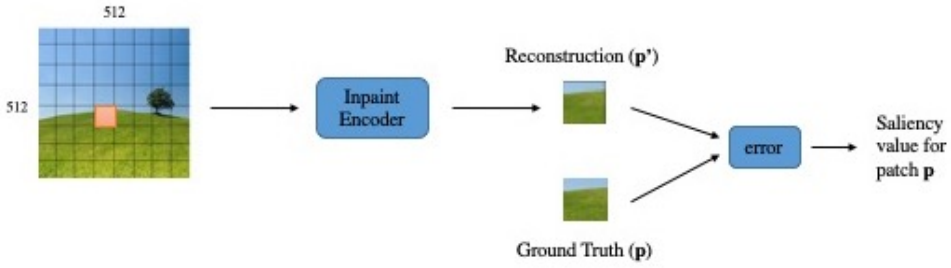


Figure 3.1: Simple diagram of our model

This simple architecture has a lot of desirable properties. First, it implicitly combines global and local contrasts by letting the network reason across the whole image and choose weights for these different aspects accordingly, as opposed to [12] that merely adds them up. Second, if the training is performed with a sufficiently varied data set, the network incorporates the prior statistical redundancies found in these images to reconstruct patches for future images. The rarity of a patch is then a combination of both its dissimilarity with the surrounding areas and with the range of natural images in general. Finally, the model doesn't require eye fixation data, working in a purely unsupervised manner.

In [119], they also use a reconstruction-based approached using an auto-encoder. However, in this case, the input is simply a patch extracted from the image and the output is a smaller patch representing the center of the input. The training is solely done by extracting random patches from the current image. Since foreground regions present a smaller probability of being sampled than background regions, the reconstruction error will be greater in those locations. In this way, the model combines the idea of CS contrast with an implicit consideration for the global aspect of the image through random sampling during training. Patches are extracted sequentially to cover the whole image and a final saliency map is generated.

This strategy seems nonetheless pretty *ad hoc*. The parameters of the auto-encoder

should represent a prior information that incorporates a wide range of natural stimuli instead of just the current image. Additionally, by using small patches as input, the non-linearities of the network are limited to learn local low-level relationships between pixels. If the model seeks to extract semantic information, it needs to consider the whole image. Furthermore, the paper uses a traditional fully-connected auto-encoder, and thus doesn't leverage the advantages of a CNN.

[58] and [62] also use a reconstruction-based approach. In this case, they assume that the image boundaries describe the background and attempt to reconstruct the rest of the image using that information. The salient foreground regions should be the ones that differ the most with the boundary information and so get a higher saliency value. This solution also seems somewhat arbitrary. It is not necessarily the case that all foreground objects are contained outside the boundary region. Plus, the paper only considers low-level features.

3.2 First Implementation

3.2.1 Data

Since the model trains in an unsupervised manner, we're not restricted to using images that have ground-truth fixations. We decided to use the famous MS Coco dataset [69], since it contains a big number of complex everyday scenes with a wide array of different objects and situations. We trained our model with 118,000 images and used 5,000 images for validation. Our first model compressed the images to 128x128 to make training easier.

For testing, we use the dataset provided in [30]. It consists of 1003 images of various scenes along with eye tracking data provided by fifteen users who free-viewed the images. We then use the measures presented in section 3.6.1 to check the effectiveness of the model and we compare it to other existing solutions in section 3.6.

3.2.2 Architecture

A lot of architectural decisions have to be made for an implementation of the aforementioned model. In [86], they perform a similar computation, where the goal of the

auto-encoder is to inpaint random missing regions. The results are very good and they show that the network is able to learn robust, generalizable features. For this reason, we implement an architecture that is very similar to the one they present. Specifically, the encoder consists of six layers of convolutions applied with 4x4 kernels, a stride of 2 and a padding of 1. This means that after every convolution, the feature map dimensions are divided by two. The last convolutional layer is applied to 4x4 feature maps, so the resulting representation before decoding is a set of 4000 1x1 features. In this way, relationships between all the regions of the image can be computed without the need of a fully-connected layer. The decoder has a similar architecture, where convolutions are replaced by transposed convolutions [26], in order to increase the dimensionality of the feature maps to the original size of the image. In order to introduce non-linear relationships between the original image and its high-level representation, we append a ReLU function [37] after every convolution. For the final one, we replace it with a Tanh function in order to limit the range of values from from -1 to 1, given that we pre-processed the images to have values falling in that range.

For training, we apply a random mask that sets four possibly overlapping square regions to zero. The network takes this as input and generates a reconstruction with the same dimensions. An Euclidean loss is then applied with a masking that ignores errors that sit outside the missing regions:

$$\mathcal{L}_{rec}(x) = \|M \odot (x - F((1 - \mathbf{M}) \odot x))\|_2^2 \tag{3.1}$$

We set the mask $\mathbf{M}$ to include the surrounding pixels of the missing regions and accentuate the loss of these surrounding regions. This encourages the network to use the context for the reconstruction. The size of the surrounding area and its level of accentuation (by multiplying the loss of that area) were left as hyperparameters that were tuned.

In order to update our parameters, we use an Adam optimizer as opposed to regular SGD, since it speeds up the convergence of the model [57].

Problem: The simple loss function corresponding to the L2-distance is able to produce a rough outline of the predicted region, but fails to capture any high frequency details. This is because such a loss is encouraged to give blurry results in order to minimize the mean pixel error. Furthermore, the network didn't seem to be able to give consistent results that revealed strong semantic understanding. Before we use it to produce

saliency maps, we need to implement a better inpainting encoder.

3.3 Second Implementation

3.3.1 Architecture

For this new implementation, we seek to use an inpainting encoder that generates more meaningful reconstructions. Luckily, [70] proposes a model that does just that. They provide very promising results with reconstructions of randomly generated masks that cover up to 60% of the image.

Figure 3.2: Reconstruction example, taken from [70]

In order to accomplish this, they introduce a series of techniques that target problems specific to the task of inpainting:

1. Instead of regular convolutions, the model uses what they call *Partial Convolutional Layers*. Let $\mathbf{W}$ refer to the convolution filter weights, $\mathbf{U}$ be the input for the current convolutional layer and $\mathbf{M}$ be the corresponding binary mask. The partial convolution at every location is then defined as:

$$u' = \begin{cases} \mathbf{W}^T(\mathbf{U} \odot \mathbf{M}) \frac{\text{sum}(\mathbf{1})}{\text{sum}(\mathbf{M})} & \text{if sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

   where $\odot$ denotes element-wise multiplication, and $\mathbf{1}$ has the same shape as $\mathbf{M}$ but every element equal to 1. This method makes it so that output values only

depend on the unmasked inputs. After each convolution, the mask is updated in the following way:

$$m' = \begin{cases} 1, & \text{if sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

meaning that if the convolution is able to condition its output on at least one valid input, the location becomes valid.

2. The loss function used is a sum of different losses that are meant to quantify different aspects of the inpainting quality. The first two terms refer to the pixel-wise differences between the reconstruction and the original image, where the first term $\mathcal{L}_{\text{hole}}$ refers to the reconstruction in the masked regions, while the second term $\mathcal{L}_{\text{valid}}$ computes the difference in the rest of the image. This is similar to the loss used in our previous architecture but, contrary to our implementation, it always includes the whole image and gives a higher weight to the currently masked region. There are also two other terms, $\mathcal{L}_{\text{perceptual}}$ and $\mathcal{L}_{\text{style}}$. The former was first introduced by [34] and computes the difference between some high-level features extracted from the original image and the reconstruction. In order to obtain these features, the images are fed into a pre-trained deep convolutional network [101]. Finally, a final loss term ($\mathcal{L}_{\text{TV}}$ is added to compute the difference between the pixels at the border of the mask and the ones just outside of it, in order to promote a smooth transition. In the end, the final combined loss is a weighted sum of these losses.

3. The encoder and the decoder are based on the famous U-Net architecture [99], where feature maps from the encoding stages are concatenated with the ones generated by the decoding module at every layer.

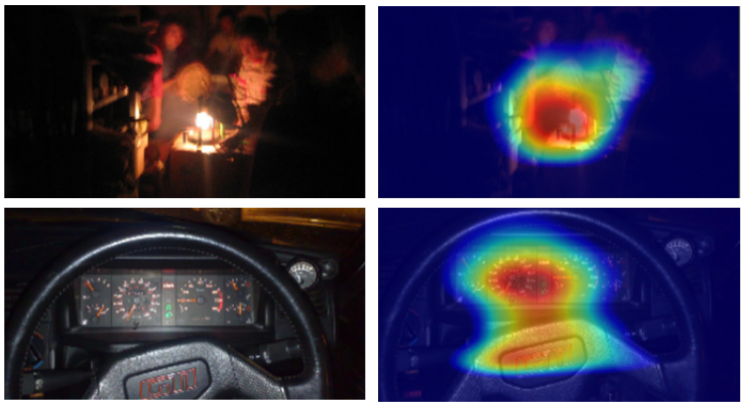
We base our implementation on the Keras version provided in [40]. It provides a reconstruction model that is pre-trained on ImageNet [101], an enormous dataset consisting of more than one million images. The images in this dataset are nevertheless simpler than the ones provided by MS COCO, since they aim to be classified under one single class; there is often a prominent object placed around the center of the image. For this reason, we train the network further with our MS COCO dataset in order to introduce the network to more complicated scenarios where interesting features are spread across the image. In the end, our reconstruction model is very powerful and is able

to use high-level semantic cues in order to fill the gaps. As opposed to our previous architecture that compresses the images to 128x128, this one re-sizes them to 512x512.

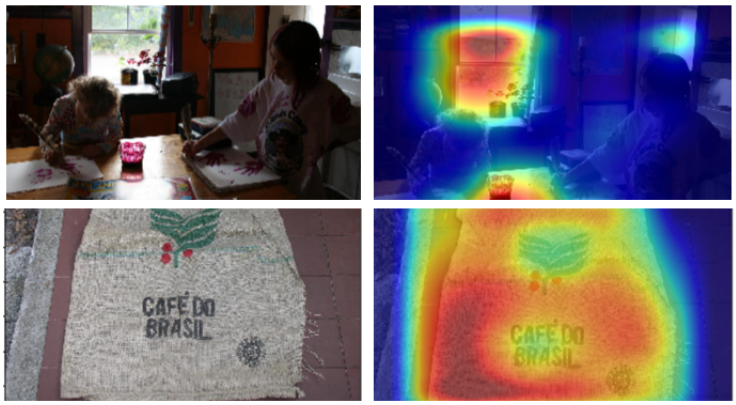
For each patch of the image, we compute its reconstruction, measure its error against the ground-truth patch and assign a saliency value to that region. We do this for every patch of the image and end up with a saliency map. We set the patch size to be 96x96, as smaller patches seemed to be trivial for the model and bigger ones severely deteriorate the results. We convolve the patch region with strides of 48, so that we end up with a final 9x9 saliency map, in which each pixel represents the error value. The map is then re-sized to the original image to yield the final saliency map.

3.3.2 Qualitative Results

From the results, we can see that the model is still lacking. While it is able to identify the region of interest for simple cases where strong low-level contrasts are present, as in the examples shown figure 3.3 (a), it still does not reason semantically like we want. In figure 3.3 (b), the first saliency map focuses on the luminance contrast by the window and ignores the presence of people. The second one focuses on the rugged texture of the material and not in the text.



(a)



(b)

Figure 3.3: Examples of saliency maps obtained with the second implementation, where the original image is on the left and the saliency map is on the right. Images in (a) show good results, while images in (b) are problematic

Problem: The model still doesn't account for higher-level cues and performs in a fashion similar to the early models of saliency.

3.4 Third Implementation

3.4.1 Architecture

In order to account for this problem, we need to rethink the way in which we compute the error between the reconstruction and the original patch. It is clear that a pixel-wise

difference is not a good measure to estimate the distance between the two patches. For example, a reconstruction that is able to recreate a highly complex object but shifts the values one pixel to some direction, could lead to a higher error than a reconstruction that merely outputs the mean intensity of the pixels. If the reconstruction model is able to correctly infer the presence of an object in the missing region, it means that the model expects that object in the scene, implying that the level of surprise should be low, and thus the error should be low. Take the example of Figure 2.7. If there was an artistic installation instead of a tree, it seems clear that the level of surprise should be higher, as a tree is a much more plausible candidate. These differences in semantic information can never be encapsulated by an L1-distance function. Ideally, we want an error function that is able to measure the perceptual difference between the two patches in a way that includes differences in semantic content.

We solve this by passing both the image with a reconstructed patch and the original image through a convolutional encoder that transforms the raw data into high-level features. After this, we can apply a simple difference as before between these new sets of features. Now, this L1-distance doesn't refer simply to pixels but to high-level latent information. Our new model can be visualized in a simple manner like this:

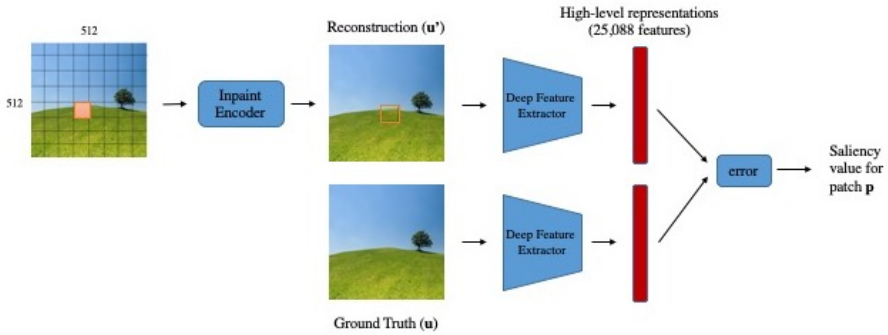


Figure 3.4: Diagram of our improved model from the third implementation

This time, we quantify saliency as the change in the representation of the entire image when a given patch is reconstructed based on its context. Now, a cluttered and chaotic background that can lead to high reconstruction errors doesn't equate to high saliency, since the high-level representation of the image in general doesn't change much.

This idea seems good, but we're still missing a way to implement such a solution. To carry this, we need a network that is trained to yield generalizable, robust and diverse features from image data. Since this insight is similar to the one proposed in the loss

function of [70], we employ the same strategy of using a VGG-16 network [106] pre-trained on classifying the ImageNet dataset mentioned previously. This dataset has 1000 classes, a plethora of different objects, and it has been shown that the features learned are generalizable enough to be applied to a wide array of different tasks [123]. It is important to note that using such an approach doesn't compromise our initial efforts of attempting to create an unsupervised model for saliency. Even though the feature extractor we're using was trained in a supervised way, we could just as easily use a trained unsupervised network. Additionally, our model is predicting eye fixation data and, in that context, it is working in an unsupervised manner as it never sees fixation data during training.

3.4.2 Qualitative Results

This time, the saliency maps of the images that gave problems in the previous implementation are much better, since the model is able to identify the presence of people and text, which are elements that strongly attract one's gaze (Figure 3.5).

The model is also able to identify high-level features such as cars, faces of animals and people in a cluttered scene (Figure 3.6). Some images still give problems though. For example, in Figure 3.7 (b), the model fixates on a sign with text, and ignores the stark contrast in luminance at the center of the image.

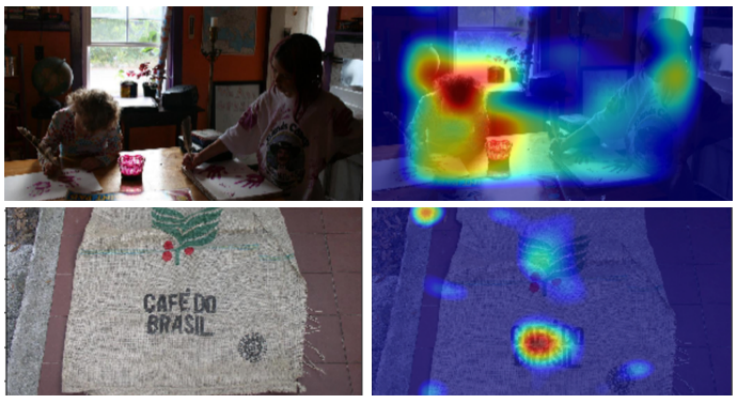
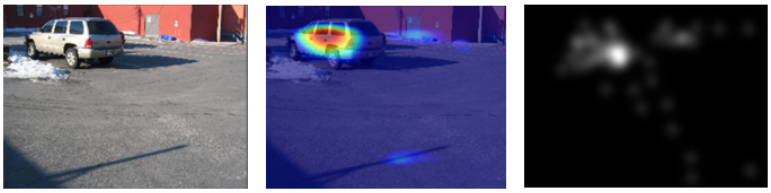
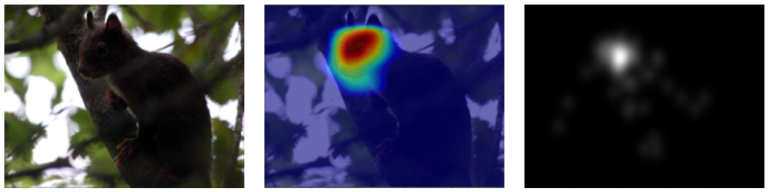


Figure 3.5: Examples of saliency maps obtained with the second implementation, where the original image is on the left and the saliency map is on the right. Both images show better results than before



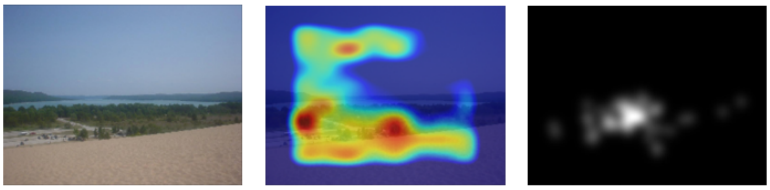
(a)



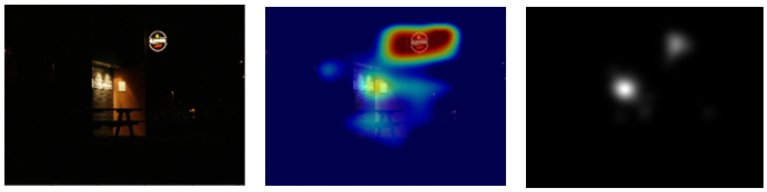
(b)

(c)

Figure 3.6: Different examples where the original image is on the left, the saliency map produced by the model is in the middle and the ground-truth fixation map is on the right



(a)



(b)

Figure 3.7: Different examples where the original image is on the left, the saliency map produced by the model is in the middle and the ground-truth fixation map is on the right

Problem: This model still faces some problems when the image doesn't present any salient semantic objects, like a wide landscape. This is due to the fact that for such images, the salient regions are defined by low-level contrasts which are not explicitly present in the features extracted by the deep network.

3.5 Final Model

3.5.1 Architecture

The solution to this new problem is a simple one. Now, instead of simply extracting the features from the last layer of the VGG-16 network, we collect the output after each convolutional block. In this way, low-level features from the first layers are combined with higher-level ones from the last layers. The final diagram for our model is the one presented in Figure 3.8:

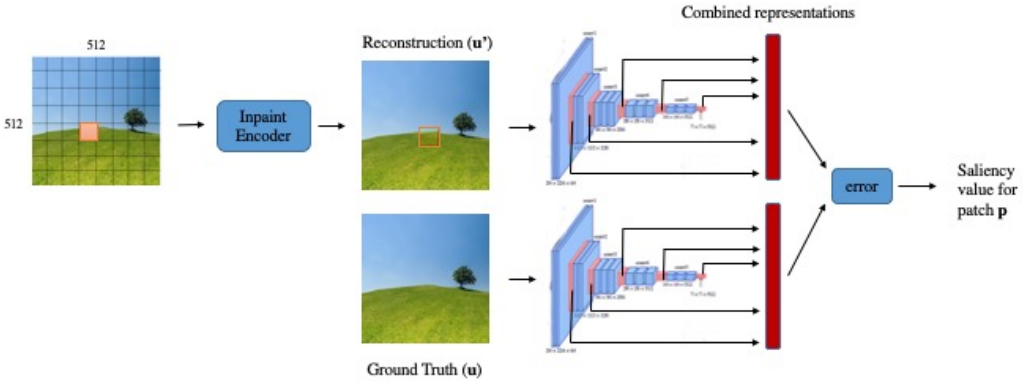
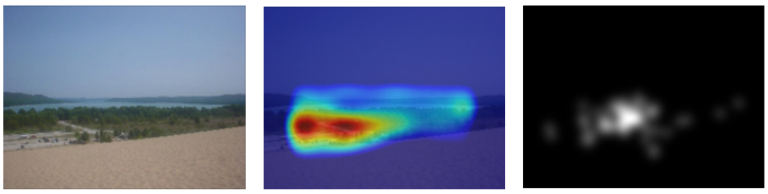


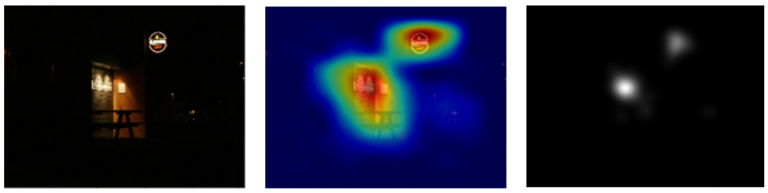
Figure 3.8: Diagram of our Final Model

3.5.2 Qualitative Results

This time, the errors the model was making in Figure 3.7 are fixed. The low-level contrasts are now taken into consideration (Figure 3.9).



(a)



(b)

Figure 3.9: Different examples where the original image is on the left, the saliency map produced by the model is in the middle and the ground-truth fixation map is on the right

3.6 Comparative Results

All the analysis we have made so far was based on a qualitative analysis of the results. We need a more tangible measure to evaluate the difference in performance for each iteration of our implementations and to compare our solution with already existing ones.

3.6.1 Evaluation Measures

So we have a working model that generates saliency maps S and we wish to compare them to ground-truth fixation maps F. It is not as trivial as one would think and there is still a debate about the best way to measure the efficiency of a saliency model. For a detailed discussion of the trade-offs between different evaluation measures, we refer the reader to [19] [97]. For the purpose of this work, we limit ourselves to explaining four commonly used methods:

- **The Kullback-Leibler (KL) divergence** is a distribution-based metric and is normally used to estimate the dissimilarity between two distributions. However, it can also be used in this case by considering both S and F as probability distributions. In this sense, KL measures the information that is lost when S is used

to approximate F:

$$KL_{div} = \sum_{x=1}^{X} F(x) * \log\left(\frac{F(x)}{S(x)+\varepsilon} + \varepsilon\right) \tag{3.4}$$

where X is the number of pixels, and the maps are normalized to have their pixel intensities sum to 1. One advantage of this measure is that it is invariant to reparameterizations, meaning it doesn't care about the absolute values of the intensities and focuses on their relative distribution.

- **The Normalized scanpath saliency (NSS)** [87] is a value-based metric that selects the saliency map's values at the eye fixation locations and normalizes them with the saliency map variance:

$$NSS(p) = \frac{S(p) - \mu_S}{\sigma_S} \tag{3.5}$$

where p is the location of one fixation and S is normalized. This measure seeks to find a balance between recall (as the score gets higher if the values are high at fixation locations) and precision (as the score is decreased if there is a small difference between fixations values and all other values). The score for each fixation is then summed and divided by the number of fixations.

- **The similarity metric (SIM)** is another distribution-based metric and it calculates the sum of the minimum value at each point in the two normalized distributions:

$$S = \sum_{x=1}^{X} \min(S(x), F(x)) \tag{3.6}$$

A similarity score of 1 occurs when both distributions are identical, and is equal to 0 when they have no overlap.

- **The linear correlation coefficient (CC)**, also named Pearson correlation coefficient, ranges from -1 to 1, where values in both extremes indicate a strong correlation between the maps:

$$CC = \frac{\text{cov}(S,F)}{\sigma_S * \sigma_F} \tag{3.7}$$

3.6.2 Results

We decide to compare our solutions with the previously mentioned model by Itti [49] and to Bruce and Tsotsos' AIM [16]. For this matter, we downloaded an implementation of these from [55] and [17] respectively.

We also include the performance of our model when extracting only high-level features (our implementation from section 3.4), only low-level features (by solely considering the output of the first layer of our feature extractor) and when it combines them into one representation (our final model).

Additionally, we have already shown the importance of adding a feature extractor before computing the error (from sections 3.3 to 3.4). However, we haven't measured the importance of our reconstruction network in our final model. In order to do so, we also run the tests on a model in which the reconstruction is replaced by a black square.

We use the four evaluation measures presented in section 3.6.1 with the test set described in section 3.2.1. We obtain the following results:

Model	**KL**	**NSS**	**SIM**	**CC**
AIM	1.7637	0.4656	0.2736	0.2403
Itti	1.7789	0.3478	0.2660	0.1912
High	1.6835	0.8115	0.3478	0.3645
Low	1.7283	0.8200	0.3471	0.3549
Final Model	**1.6427**	**0.87**	**0.3511**	**0.3832**
No Rec.	1.7345	0.7528	0.3453	0.3403

Table 3.1

Finally, to make sure that our model is able to reason in a global way, we test it against synthetic images that present a symmetrical texture and we compare it with Itti's model:

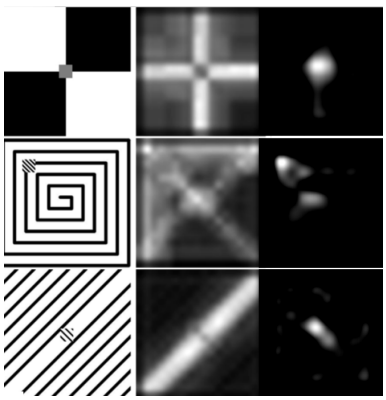


Figure 3.10: Left: Original Image, Centre: Itti's Maps, Right: Our Maps

3.7 Conclusion and Future Work

3.7.1 Conclusion

Throughout this dissertation, we sought to understand some fundamental ideas underlying neural information processing. From the insights that we uncovered, we defined attention in free-viewing situations under the umbrella of predictive coding, where gaze is directed towards the locations which create the most surprise, or that lead to the highest prediction error. We transformed this idea into a working implementation, by equating saliency to the reconstruction residual of trying to generate a missing region from its surrounding context. We finally showed that this model leads to good results, by being able to incorporate the success of deep learning with a reconstruction-based approach inspired by predictive coding.

3.7.2 Limitations and Future Work

Despite the decent results, our model still has a lot of limitations that can be further improved:

1. Arguably the most pressing one is that the model is fully deterministic. The ideas of a Bayesian brain or of a free-energy principle are fundamentally probabilistic. Our model highlights prediction errors but doesn't explicitly reason with uncertainty. Such a problem seems to be much harder to tackle. A possible avenue would be to explore the use of a variational auto-encoder [58], in which the output of the network is probabilistic, but such a recommendation remains purely speculative

2. Another limitation of our model is that it merely reasons with spatial predictions, while ignoring any temporal information. It's easy to see that images are highly correlated in time, and a brain that is actively engaging with the external world certainly takes advantage of such dependencies. Many models of representation learning already exits that use this temporal information to predict future states [81]. It would be interesting to extend our model to video data, where prediction would also occur in the temporal axis.

3. One other problem that seems worthy of mention is that we were constrained by

the reconstruction model's architecture to work with 512x512 images and that the patch stride of 48 can make the model lose some information. Smaller strides could be explored for more precise saliency maps.

4. The fact that we have two eyes allows us to perceive depth, and it is estimated that this aspect has a large impact in our perception of saliency [22]. With the advent of large stereoscopic datasets, it would be interesting to explore the construction of a model based on ideas of predictive coding applied to such datasets.

Bibliography

[1] Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. In *International conference on computer vision systems*, pages 66–75. Springer, 2008.

[2] J Yu Angela and Peter Dayan. Inference, attention, and decision in a bayesian neural architecture. In *Advances in neural information processing systems*, pages 1577–1584, 2005.

[3] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

[4] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. In *ACM Transactions on graphics (TOG)*, volume 26, page 10. ACM, 2007.

[5] Edward Awh, Artem V Belopolsky, and Jan Theeuwes. Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in cognitive sciences*, 16(8):437–443, 2012.

[6] William F Bacon and Howard E Egeth. Overriding stimulus-driven attentional capture. *Perception & psychophysics*, 55(5):485–496, 1994.

[7] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

[8] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[9] Pietro Berkes, Gergő Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.

[10] Rafal Bogacz. A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, 76:198–211, 2017.

[11] Ali Borji. Saliency prediction in the deep learning era: An empirical investigation. *arXiv preprint arXiv:1810.03716*, 2018.

[12] Ali Borji and Laurent Itti. Exploiting local and global patch rarities for saliency detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 478–485. IEEE, 2012.

[13] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013.

[14] Ali Borji, Dicky N Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013.

[15] Ali Borji, Hamed R Tavakoli, Dicky N Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 921–928, 2013.

[16] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006.

[17] Neil D. B. Bruce. Aim. https://github.com/TsotsosLab/AIM, April 2009.

[18] Neil DB Bruce and John K Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):5–5, 2009.

[19] Neil DB Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K Tsotsos. On computational modeling of visual saliency: Examining whats right, and whats left. *Vision research*, 116:95–112, 2015.

[20] Robin L Carhart-Harris and Karl J Friston. The default-mode, ego-functions and free-energy: a neurobiological account of freudian ideas. *Brain*, 133(4):1265–1283, 2010.

[21] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6:27755, 2016.

[22] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters*, 23(6):819–823, 2016.

[23] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Multi-level net: A visual saliency prediction model. In *European Conference on Computer Vision*, pages 302–315. Springer, 2016.

[24] Stephen V David, Benjamin Y Hayden, and Jack L Gallant. Spectral receptive field properties explain shape selectivity in area v4. *Journal of neurophysiology*, 96(6):3492–3505, 2006.

[25] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[26] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[27] Wolfgang EinhÃ, Ueli Rutishauser, Christof Koch, et al. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of vision*, 8(2):2–2, 2008.

[28] Harriet Feldman and Karl Friston. Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215, 2010.

[29] Society for Neuroscience. *Brain Facts*. 2018.

[30] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

[31] Karl Friston. Active inference and free energy. *Behavioral and Brain Sciences*, 36(3):212–213, 2013.

[32] Karl Friston. Consciousness and hierarchical inference. *Neuropsychoanalysis*, 15(1):38–42, 2013.

[33] Karl Friston. Does predictive coding have a future? *Nature neuroscience*, 21(8):1019, 2018.

[34] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[35] HJM Gerrits, B De Haan, and AJH Vendrik. Experiments with retinal stabilized images. relations between the observations and neural data. *Vision research*, 6(7-8):427–440, 1966.

[36] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[37] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[38] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[40] Mathias Gruber. Partial convolutions for image inpainting using keras. https://github.com/MathiasGruber/PConv-Keras, 2018.

[41] Tim Halverson and Anthony J Hornof. A minimal model for predicting visual search in human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 431–434. ACM, 2007.

[42] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[43] John M Henderson and Andrew Hollingworth. High-level scene perception. *Annual review of psychology*, 50(1):243–271, 1999.

[44] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Ieee, 2007.

[45] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015.

[46] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[47] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[48] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194, 2001.

[49] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.

[50] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012.

[51] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[52] Timor Kadir and Michael Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[53] Timor Kadir and Michael Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[54] J Kevin O'Regan, Heiner Deubel, James J Clark, and Ronald A Rensink. Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, 7(1-3):191–211, 2000.

[55] Akisato Kimura. pysaliencymap. https://github.com/akisato-/pySaliencyMap, 2014.

[56] RA Kinchla and JM Wolfe. The order of visual processing:top-down,bottom-up, or middle-out. *Perception & psychophysics*, 25(3):225–231, 1979.

[57] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[58] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[59] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[60] Kathryn Koehler, Fei Guo, Sheng Zhang, and Miguel P Eckstein. What do saliency models predict? *Journal of vision*, 14(3):14–14, 2014.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[62] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017.

[63] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[64] Matthias Kümmerer, Thomas Wallis, and Matthias Bethge. How close are we to understanding image-based saliency? *arXiv preprint arXiv:1409.7686*, 2014.

[65] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017.

[66] Quoc V Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S Corrado, Jeff Dean, and Andrew Y Ng. Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:1112.6209*, 2011.

[67] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[68] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2976–2983, 2013.

[69] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[70] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[71] Norman H Mackworth and Anthony J Morandi. The gaze selects informative details within pictures. *Perception & psychophysics*, 2(11):547–552, 1967.

[72] Christopher Michael Masciocchi, Stefan Mihalas, Derrick Parkhurst, and Ernst Niebur. Everyone knows what is interesting: Salient locations which should be fixated. *Journal of vision*, 9(11):25–25, 2009.

[73] Lucia Melloni, Sara van Leeuwen, Arjen Alink, and Notger G Müller. Interaction between bottom-up saliency and top-down control: how saliency maps are created in the human brain. *Cerebral cortex*, 22(12):2943–2952, 2012.

[74] Beren Millidge. Fixational eye movements: Data augmentation for the brain? 2019.

[75] David Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 66(3):241–251, 1992.

[76] Naila Murray, Maria Vanrell, Xavier Otazu, and C Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR 2011*, pages 433–440. IEEE, 2011.

[77] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[78] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996.

[79] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[80] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

[81] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[82] Clyde W Oyster. *The human eye: structure and function*. Sinauer Associates, 1999.

[83] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[84] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

[85] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

[86] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[87] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005.

[88] Stephen Lucian Polyak. The retina. 1941.

[89] Michael I Posner, Robert D Rafal, Lisa S Choate, and Jonathan Vaughan. Inhibition of return: Neural basis and function. *Cognitive neuropsychology*, 2(3):211–228, 1985.

[90] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102, 2005.

[91] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[92] Shafin Rahman and Neil Bruce. Saliency, scale and information: Towards a unifying theory. In *Advances in Neural Information Processing Systems*, pages 2188–2196, 2015.

[93] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

[94] Robert Rauschenberger. Reentrant processing in attentional guidance–time to abandon old dichotomies. *Acta psychologica*, 135(2):109–11, 2010.

[95] Karsten Rauss and Gilles Pourtois. What is bottom-up and what is top-down in predictive coding? *Frontiers in Psychology*, 4:276, 2013.

[96] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 1153–1160, 2013.

[97] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 1153–1160, 2013.

[98] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019, 1999.

[99] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[100] David Rudrauf, Daniel Bennequin, Isabela Granic, Gregory Landini, Karl Friston, and Kenneth Williford. A mathematical model of embodied consciousness. *Journal of theoretical biology*, 428:106–131, 2017.

[101] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[102] Dario D Salvucci. An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, 1(4):201–220, 2001.

[103] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[104] Jitendra Sharma, Alessandra Angelucci, and Mriganka Sur. Induction of visual orientation modules in auditory cortex. *Nature*, 404(6780):841, 2000.

[105] Daniel J Simons and Daniel T Levin. Change blindness. *Trends in cognitive sciences*, 1(7):261–267, 1997.

[106] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[107] N Sprague and D Ballard. Eye movements for reward maximization. advances in neural information processing systems, 16. 2003.

[108] Gábor Stefanics, Piia Astikainen, and István Czigler. Visual mismatch negativity (vmmn): a prediction error signal in the visual modality. *Frontiers in human neuroscience*, 8:1074, 2015.

[109] Gabor Stefanics, Jakob Heinzle, András Attila Horváth, and Klaas Enno Stephan. Visual mismatch and predictive coding: a computational single-trial erp study. *Journal of Neuroscience*, 38(16):4020–4030, 2018.

[110] Christopher Summerfield and Tobias Egner. Expectation (and attention) in visual cognition. *Trends in cognitive sciences*, 13(9):403–409, 2009.

[111] Benjamin W Tatler, Roland J Baddeley, and Iain D Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision research*, 45(5):643–659, 2005.

[112] Jan Theeuwes. Top–down and bottom–up control of visual selection. *Acta psychologica*, 135(2):77–99, 2010.

[113] Antonio Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003.

[114] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[115] Leslie G Ungerleider. Two cortical visual systems. *Analysis of visual behavior*, pages 549–586, 1982.

[116] David C Van Essen and Charles H Anderson. Information processing strategies and pathways in the primate visual system. *An introduction to neural and electronic networks*, 2:45–76, 1995.

[117] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014.

[118] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[119] Chen Xia, Fei Qi, and Guangming Shi. Bottom–up visual saliency estimation with deep autoencoder-based sparse reconstruction. *IEEE transactions on neural networks and learning systems*, 27(6):1227–1240, 2016.

[120] Chen Xia, Fei Qi, Guangming Shi, and Pengjin Wang. Nonlocal center–surround reconstruction-based bottom-up saliency estimation. *Pattern Recognition*, 48(4):1337–1348, 2015.

[121] Chen Xia, Pengjin Wang, Fei Qi, and Guangming Shi. Nonlocal center-surround reconstruction-based bottom-up saliency estimation. In *2013 IEEE International Conference on Image Processing*, pages 206–210. IEEE, 2013.

[122] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[123] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[124] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[125] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[126] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008.

[127] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[128] Xilin Zhang, Li Zhaoping, Tiangang Zhou, and Fang Fang. Neural activities in v1 create a bottom-up saliency map. *Neuron*, 73(1):183–192, 2012.