UNIVERSITAT DE BARCELONA

Minds, Brains and Machines: Deep Learning Models of Vision

Cañas Tarrasón, Eric

Master in Artificial Intelligence

23th June 2020

Contents

1 Introduction 2

2 Mimicking the human vision system 2

2.1 Mimicking the mechanical part: How to Capture images . . . 3

2.1.1 The Eye . . . 3

2.1.2 The Camera . . . 5

2.2 Mimicking the Neural part: Understanding images . . . 6

2.2.1 The *visual cortex* and the *visual association cortex* . . . 6

2.2.1.1 The *visual cortex* . . . 7

2.2.1.2 The *visual association cortex* . . . 8

2.2.2 Deep Neural Networks . . . 9

2.2.2.1 Deep Convolutional Neural Networks . . . 10

3 Characterizing convolutional layers 10

3.1 From simple to complex representations . . . 11

3.1.1 The emergence of shared knowledge . . . 13

4 Conclusions 14

Abstract

At 1966, *Seymour Papert* failed when he tried to reproduce in a computer the human vision system as a summer project. Today, at 2020, we have a myriad of *Computer Vision* models which allows not only to describe images, but even to imagine new scenes. Since *Deep Convolutional Neural Networks* were presented in 2012, the *Computer Vision* has suffered which is maybe the higher growth rate in the *Machine Learning* field. In this work we will study how each property of the human vision system is represented or not by *Deep Learning* models, how they build their knowledge and how we can understand it as something beyond the usual black box system composed by building blocks.

1 Introduction

Since the *Last Ultimate Common Ancestor* of all living beings (*LUCA*)[1] was born, almost 4 billion years ago, evolution has been working constantly in developing the *visual system* as we know it today[2]. From our part, since 1966 summer[3] until nowadays, humans have teach our computers to mimic this behaviour in less than a 10^{-8} part of the time. This process, obviously, has neither been direct nor simple, but has relied on a large set of technologies which range from cameras, for mimicking the complex mechanical system of the eye, until *Deep Artificial Neural Network* models, for replying the most basic characteristics of the neural part of the vision system.

When speaking about this neural system replication, there is without any doubt a concrete point of time which must be highlighted as the starting point of the *Computer Vision* spring: December of 2012[4]. It is the point, where three researchers from the university of Toronto: Hinton, Sutskever and Krizhevsky, won the most important challenge on computer vision classification, *ImageNet*[5], and changed the paradigm of the *Computer Vision* field for the rest of times, by including to the usual *Neural Network* scheme the concept of *spatial information.* Since this point, *Computer Vision* has not only beaten humans in some classification tasks[6][7], but also have given astonishing results in more complex fields like describing scenes[8] or even generating new ones[9].

In this work, we will study how the basic scheme of human vision is replicated by these *Deep Learning* models, and how knowledge emerge from them. For this purpose, we will focus first on how images are captured, and then, on how they are processed and transformed into knowledge by both, our *visual system* and *Deep Convolutional Neural Networks.*

2 Mimicking the human vision system

The visual system is the part of the nervous systems that captures the light waves (in the visible spectrum) that comes into our eyes and transform them into understandable information about the environment around us. This process starts into the eye, where light enters by the *cornea* and passing through the *pupil* arrives to the *retina.* At this point, it is transformed into a nervous pulse which will end by reaching the brain through the *optic nerve.* Here is were process of capturing images is performed. Since this point onward, the nervous pulse will travel through the *optic chiasma* and the *optic tract* until the *optic tectum*, that is the part where most vertebrates as reptiles, fishes or birds focuses their vision processes [10]. However, in more complex mammals like humans, these pulses continues until the *occipital lobe*, in the back of the brain, where they are processed by the *visual cortex* and the *visual associtive cortex.* Here is where the process of understanding the visual information is performed.

In this section, we will deeply analyze how these two main processes (capturing and understanding) are mimicked by the current human technology.

2.1 Mimicking the mechanical part: How to Capture images

At 1604 the term of *camera obscura* was introduced by first time by Johannes Keppler on his treatise: *Ad Vitellionem Paralipomena*[11]. This term referred to an antique instrument (firstly described in VI a.C) that, capturing the light through a pinhole (like *pupil* in the *eye*), projected in a surface (like the back of the *eye*) a (reversed) image of what is in front of it (figure 1 (a)). This *pinhole* could include or not a *diaphragm* mechanism for regularizing the amount of light which is captured (like *iris*). However, it was not until 1826 that first permanent photography was taken by the chemist *Joseph Nicéphore Niépce* (figure 1 (b)) projecting these light into a photo-sensible material (like *retina*). Finally, in 1961 the physical *James Clark Maxwell* took the first color image (figure 1 (c)) demonstrating that all colors could be captured by passing the light through different filters, dyed with three primary colors: *Red*, *Green* and *Blue* (somehow like *retina cons*).

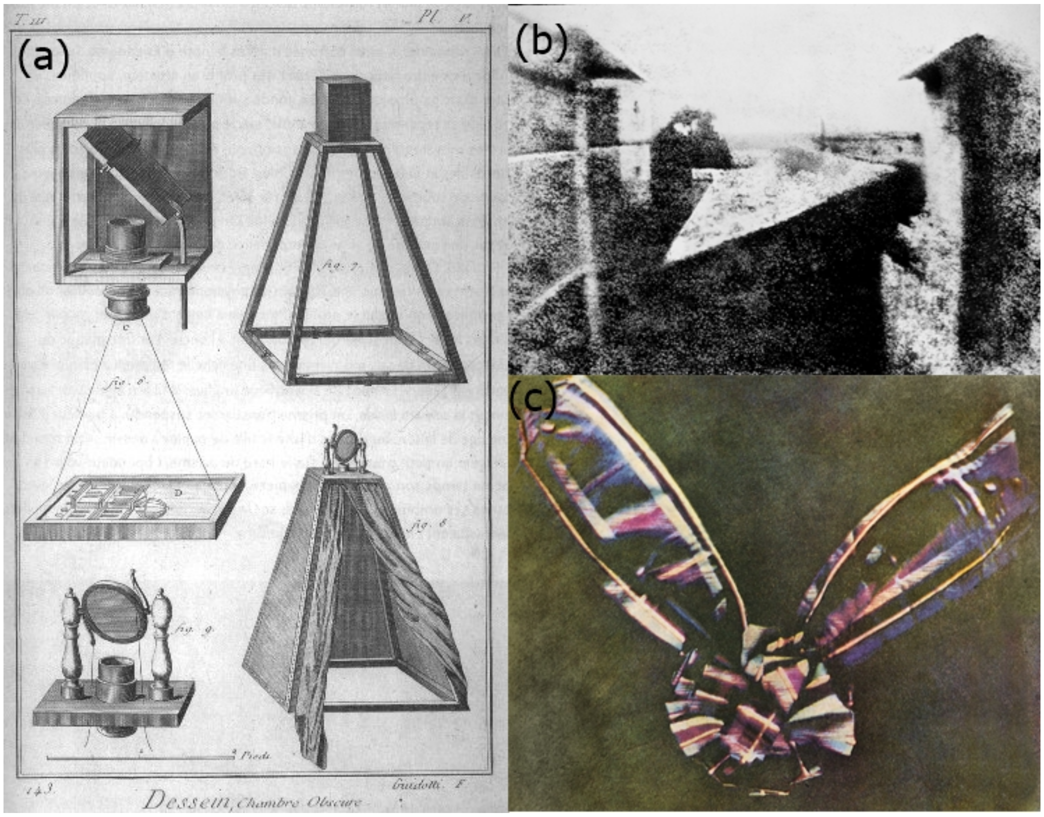

Figure 1: Evolution of the photography. (a): Plains of XVIII century detailing the projection system of the *camera obscura*. (b): *View from the Window at Le Gras*, first permanent photography took by *Joseph Nicéphore Niépce*. (c): First color photography, took by *James Clark Maxwell*.

In this section we will analyze how the process of taking a photograph mimics the process that the *human eye* performs for capturing images.

2.1.1 The Eye

The *human eye* is a complex photo-receptor organ that can be understood from several point of views, such as *chemical*, *psychical*, *medical*, etc. For the sake of simplicity, here, we will define a basic scheme of how the *eye* works from the *functional* point of view, in order to make it comparable with the systems that concerns us, the *Computer Vision* technologies.

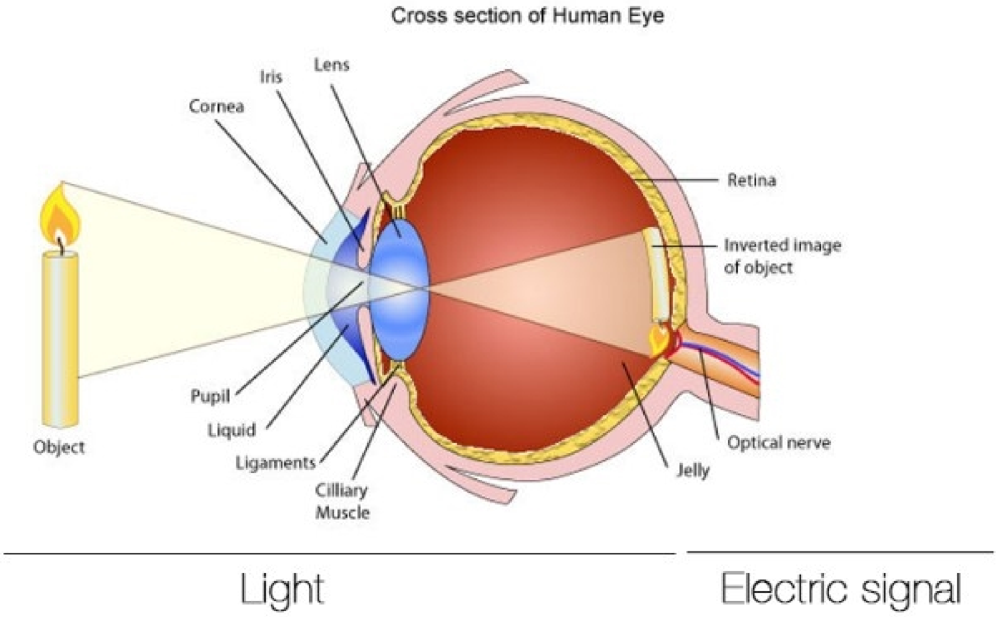

Figure 2: Basic schema of how the *human eye* deals with images[12]

1. The light waves reflected by the object arrives to the *cornea*.
2. In the *cornea*, those light waves are refracted. In this way, parallel rays are re-directed for preparing its convergence.
3. Those rays pass through the *pupil*, which is a *pinhole* whose diameter is managed by the *iris* (an *sphincter* muscle which works as *diaphragm*). The diameter of the *pupil* determines which is the amount of light which is captured. It is used for reducing or enhancing the luminosity according to the luminous situation.
4. The light arrives to the *lens*, that is in charge of managing the point where the light converges, in order to manipulate the distance at which we focus (the *focal distance*). Its form is managed by the *ligaments*.
5. After convergence, light arrives to the *retina*, a tissue completely covered by photo-receptors that is able to translate the light into neural electric pulses, that could be interpreted by the brain.
6. Those electric pulses are transmitted through the *optical nerve*.

Here, we have shown how the light is manipulated, in order to prepare its correct reception. However, it is important to focus especially in the *retina* the part where it is translated understandable information.

Retina is composed by several layers of photo-receptor cells. However, as we are focusing on the *functional* aspects of the vision we will focus on a specific one, the *nuclear layer*. This layer contains two main kinds of cell bodies which work as photo-receptors:

1. **Rod Granules:** They are the most numerous kind of cells in this layer, and are in charge of the *scotopic vision*. They can not detect any *color*, give *low resolution* vision and are activated by *low light levels*. They are responsible of the *rapid movement detection*, the *vision in the darkness* and the *peripheral vision*.

2. **Cones Granules:** They are in charge of the *photopic vision* (related with colors). They are much less numerous than *rods* and are mostly concentrated in the center of the *retina* (the *fovea*). They are only activated when the level of light is high enough. There are three kind of *cones* (*Red* (about 64%), *Green* (about 32%) and *Blue* (about 4%)) and each one maximizes its reaction on receiving a concrete *wavelength*. (figure 3). They are responsible of the *high resolution vision*, where we can perceive *accurate details* and *colors*.

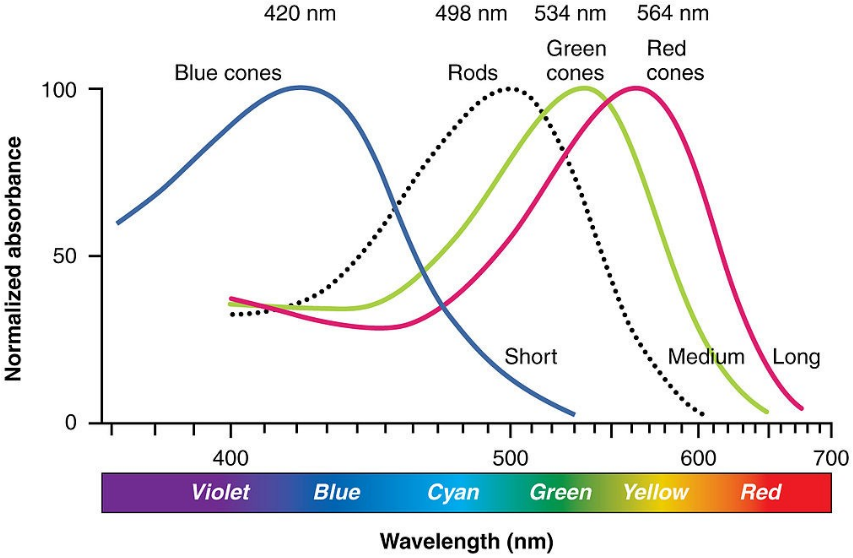

Figure 3: Normalized absorbance of each one of the photo-receptors cell bodies in the nuclear layer. Three kind of *cones* are represented by the color to which they most react. *Rods* are represented by the dashed curve.

Once we have defined the main process through which the eye is able to capture images and transform them into nervous electric pulses (understandable by the brain), lets see how *Computer Vision* mimics this process for obtaining its understandable input data, the images.

2.1.2 The Camera

As well as *eyes* transform the light into neural electric pulses, that are understood by the brain, modern *photo camera* transforms the *luminous temporal information* into persistent electrical information, that can be understood by computers.

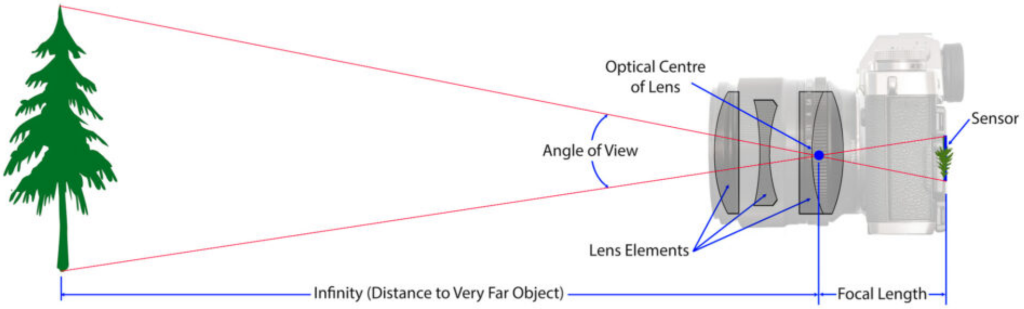

Figure 4: Simplified scheme of the capturing process that a digital camera performs. Compare it with figure 2.

So here, we will define how each part of the aforementioned *human capturing process* is interpreted and represented by the process of taking a digital picture.

1. The light waves reflected by the object arrives to the first *lens*.
2. In this *lens*, light waves are refracted like in *cornea* in order to prepare them to converge at the last *lens* (equivalent to the *eye lens*).
3. Those rays pass through the *pinhole* (equivalent to the *pupil*) whose diameter is managed by the *diaphragm*, that works like the *iris*. This aperture controls the amount of light which will enter in the *sensor*.
4. The light rays passes through a set of *lenses* which are in charge of managing its convergence. They exactly actuate like the *eye lens*, but here the *focal distance* is not managed by their form, but by the distance between them.
5. After convergence, the light arrives (reversed) to the *sensor* (equivalent to the *retina*) which is made of photo-receptors which will transform the luminous information to electric pulses.
6. Those electric pulses are stored digitally.

Since here, the process of manipulating light has been quite similar to the one performed by the *human eye*. However, when we focus in the internal similarities between the *retina* and the *sensor* we can find more differences.

In camera *sensors*, we do not find anything as *rodes* or *cones*. All digital cells have the same sensitivity (which is parameterizable) and are distributed in the same proportion along the *sensor*. Color information is also captured by *Blue*, *Green* and *Red* specialized cells (like *cones*), but the proportions change. The minimum visual unity, the *pixel*, is composed by a 2x2 square of cells: A *Blue* cell, a *Red* cell and two *Green* cells. Two *Green* cells are included for completing the square, since *Green* is the color which is more in the center of the spectrum among those three.

2.2 Mimicking the Neural part: Understanding images

Since this point, we have seen how the first part of the *visual system* has been satisfactorily replied by humans. This was, in fact, the easy part, since the system of the *eye* is well known. However, it is not the same for the brain, the most unknown organ of the human body, whose systems are orders of magnitude more complex. In this section we will focus on defining the main parts of this *visual cortex* and the *visual association cortex*, in order to compare them with the most similar computational framework that humans have designed: The *Deep Convolutional Networks*.

2.2.1 The *visual cortex* and the *visual association cortex*

When the *neural electric pulse* is delivered by the *retina*, it travels through *theoptic chiasma*, the *theoptic tract* and the *theoptic tectum* before arriving to the *visual cortex*, that is the part where most of these pulses are transformed into visual knowledge, before being transformed into general knowledge into the *visual association cortex*. There are several parts of the brain which are related with the vision and explaining all of them would be neither possible nor useful for the purpose of this work, since we are not studying the brain, but the *deep learning* systems which mimic its *visual function*. By this reason and for the sake of simplicity, we will focus only on the pair *visual cortex-visual association cortex*, due to its especial relevance. However, is important to denote, these systems are extremely complicated and several details about them remain unknown, therefore, in this section we will only depict which are the main characteristics that are known about each one of their most relevant parts.

2.2.1.1 The *visual cortex*

The *visual cortex* presents a hierarchical structure regarding their neurons connectivity (although there are tons of shortcuts among them). It is composed by five main regions: *V1* (*primary visual cortex*), *V2* (*secondary visual cortex*), *V3*, *V4* and *V5* (*middle temporal visual area* (*MT*)) (figure 5).

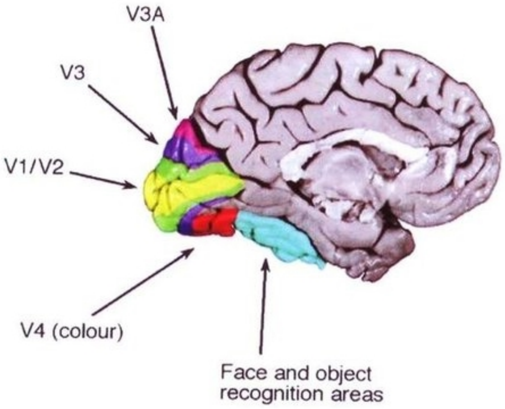

Figure 5: Simplified scheme of the main regions which composes the *visual cortex* and part of the *visual associative cortex*[13]. *Middle temporal visual area* (*V5*) is not shown because, in lateral view, it is located below *V4* region.

***V1* (*primary visual cortex*):** This is the entrance point of the *signal* and is the region where lower abstraction levels are processed. It is specialized in *pattern recognition* and is also in charge of discerning between *static* and *moving objects* (without *direction* discerning). In order to simplify the received signal this part performs a process similar to *saliency maps*[14], a *computer vision* algorithm which compute pixel similarities (through histograms), in order to transform the raw image data into something more meaningful and easier to manipulate for *segmentation tasks*. It send information to the regions *V2*, *V3*, *V4* and *V5*.

***V2* (*secondary visual cortex*):** This is the entrance point of the *visual association cortex*, and the second biggest part of the *visual cortex*. Most of its neurons are specialized in detecting *colors*, *orientations*, *shapes*, *sizes* and more complex *patterns* than *V1*, based in the *spatial frequency*. However, there are also some neurons here which are able to detect *contours* and even performing *simple segmentation*[15]. Recent studies suggest that it is also relevant for the *visual memory*[16]. It receives information from *V1* and sends information to *V3*, *V4* and *V5*, as well as feedback to the *V1*.

V3: The accurate extension of this region is still controversial, however, it is clear that it receives the information from *V1* and *V2*, and have some connections with other brain parts outside the *visual cortex* and the *visual association cortex*. It is in charge of processing the *global motion* of the scene[17], as well as detecting *large patterns*. Inside the *visual cortex*, it sends information to the areas *V4* and *V5* but also sends feedback to the previous.

V4: The extension of this region remains unknown. However, it is well known that implements functionalities of *selective attention*[18], for modifying the *firing rates* of the neurons with which it is connected (in order to focus in some aspects of the scene or others). Those neurons are mostly specialized in recognizing *features* and *forms* that are more abstract than previous regions. In this aspect, it is known that it is able to recognize *geometric forms* easily, but the bounds on the complexity that it could recognize still being too fuzzy. It receives information from *V1*, *V2* and *V3* and sends information to *V5*.

***V5* (*Middle Temporal Visual Area* (*MT*)):** This region is focused on characterizing *speed* and *trajectories* of complex objects[19], as well as managing the *eye movement* for following them. It is connected with regions *V1*, *V2*, *V3*, and *V4*, as well as some regions outside the *visual cortex*.

It is important to take into account, that functionalities described here are only the ones in which most neurons of each region are focused on. However, each region includes, in different proportions, several types of neurons with different functionalities, and some of them could be overlapped between different regions. Moreover, the connections between them are quite complex and fuzzy, and each one is connected not only with every other region in stronger or weaker ways, but also with variable *fire rates* in some cases. By this reason, here, has only been described their main functionalities without entering in complex details, since the purpose of this description is not to give a deep understanding of the neuroscience surrounding the *visual cortex*, but defining a general framework for comparing it with the actual *Deep Learning* models of vision.

2.2.1.2 The *visual association cortex*

The *visual association cortex* is a vague term that refers to all these parts of the brain that are in some way related with transforming the visual information that comes from the *visual cortex* to complex general information associated with our prior knowledge[20]. By this reason, all parts of the *visual cortex* except *V1* (the *primary visual cortex*), that is the unique which entrance is not visual processed information but raw information coming from the *retina*, are considered to also be within the *visual association cortex*.

Apart of them, it is composed by several areas which processes different associative information (such as *faces recognition*, *scene recognition*, *emotional associations*...) and the processes performed by most of them still being unknown. By this reason, for the purposes of this work, only two well known of them will be depicted in order to exemplify, in general terms, how these kind of processes works.

***PPA* (*Parahippocampal Place Area*):** It is a concrete area of the brain, located in the *posterior parahippocampal gyrus* -The region of the brain especially involved in *spatial memory* and *navigation*[21]- which have a high neural response when observing (mostly, but not only) different environmental places containing: buildings, similar structures (from both, outdoor and indoor points of view) or even some related objects like chairs and other furniture[22]. In some way, it could be related with recognizing familiar places and environments that are *navigation oriented*, since damages in this region has demonstrated to severally harm the ability for keeping the orientation when navigating through familiar places[23].

***FFA* (*Fusiform Face Area*):** It is a concrete area of the brain located in the *fusiform gyrus* (blue region in figure 5), that presents high activation when observing (mostly, but not only) face images[24]. This concept of faces, in general, is most related to the identity of the subject than to the face itself, since it has been demonstrated that persons which are used to observe other identifiable objects like *birds* or *cars* also presents high neural responses in this region when observing them[25].

All these regions which are specialized on concrete visual domains are commonly developed by the usual repetitions of some patterns into the *visual cortex*. In this way, for example, neurons that are usually stimulated when observing faces forms, end by composing a specialized region that is more focused on recognize them and their finest details. Recent studies have also demonstrated that these regions are not only related with explicit visual information, since they can also present high activity when using language related with their expertise field [26].

2.2.2 Deep Neural Networks

As have been seen in previous section, *visual cortex* and *visual association cortex* are complex regions of the brain for which we have very restricted knowledge. Moreover, they are not only related with strictly *spatial information*, but also with concepts which have a *temporal* dimension like *movement*. So lets define which are the principal kind of *Deep Neural Networks* layers that could model the previous described functionalities:

***MLP* (*Multi-Layer Perceptron*):** *Multi-Layer Perceptrons* are the simplest precursors of actual *Artificial Neural Networks*[27]. They are the most used architecture when dealing with feature based problems like *regression* or *classification*. The process that they perform could be associated with the deeper parts of the *visual associative cortex* (like *FFA* or *PPA*). However, in *computer vision*, since uncorrelated *pixels* are too disperse for being considered as *features*, they have not achieved by themself great results unless they are preceded by image feature extractors like a *CNN*s.

***CNN* (*Convolutional Neural Network*):** These networks are characterized by layers which performs the convolution operator[28]. This operator relates each pixel with its neighborhood, including in the process *spatial* relations. They work specially well as image feature extractors, generating features that can be used as input for a *MLP*. The process that they perform could be compared with some parts of the *visual cortex*, that transform the raw *input* of the *retina* (image) to an abstract representation (*abstract features*) that will be later understood by different regions of *visual associative cortex*. Since 2012[4], they have been the most used networks for all kind of *computer vision* tasks, giving astonishing results that even outperforms the human abilities in some classification tasks. However, the *visual cortex* is highly related with *temporal information* like *movement* that can not be represented by simple *CNN*s.

***RNN* (*Recurrent Neural Networks*)** These networks are specialized in finding patterns inside *temporal sequences* (like *image feature sequences*)[29]. As they introduce the concept of temporal sequences, they are, in combination with *CNN*s, able mimic the functionalities of the *visual cortex* related with *movement detection* and *trajectories* definition. In the *computer vision* field they set the standard of video analysis.

***Transformers*:** *Transformers* are a recent kind of models that include in the process the concept of *selective attention* [30]. This concept is a relevant functionality of the human brain that is even present in some regions of the *visual cortex* like *V4*. In last years, *transformers* have been getting astonishing results in *Natural Language Processing* problems, however, in the *computer vision* field, although some adaptations have been proposed (as *Image Transformers*[31]), they still being mostly reserved for captioning generation problems.

As seen, *Deep Learning* defines a set of models which are able to mimic different functionalities of the *visual cortex* and the *visual associative cortex*. In this work, we will focus the attention in one of them, *Convolutional Neural Networks*, since they are not only the most used models by far and the ones that best results obtains, but also the ones from which most *interpretable information* can be extracted. For extracting information of these models we have two principal methods:

Study the activation of each neuron: That is, to study which are the patterns which maximize the activation of each individual neuron. It emulates in some way the process of the *MRI* (*Magnetic Resonance Image*), since we can obtain maps about which regions of the network are most activated when we pass a concrete image stimulus through it.

Ablation study: It consist in deleting some parts of the network and observing how the results are affected. For example, we could check if a concrete part of the network is relevant for *face recognition* purposes by ablating this part and observing how the ability of the network for recognizing faces is damaged or not. This method emulates those studies where we analyze how the abilities of a patient is damaged after loosing a concrete part of the brain.

In following sections we will study which conclusions we could extract from applying these methods, however, first, lets define the main characteristics of *Deep Convolutional Networks*.

2.2.2.1 Deep Convolutional Neural Networks

Convolutional Neural Networks have been able to outperform the human brain in some image classification tasks. By this reason they can be considered as the best opponent against the *visual cortex* and the *visual association cortex*. So lets explain which are the main characteristics that define them.

- For classification tasks they are usually composed by a set of *convolutional layers*, each one followed by an activation function and in some cases by a *pooling layer* which progressively reduce the size of the feature space. On the top of all these layers, usually a *MLP* is included in order to use the features extracted for solving the classification task. However, in newest networks, this *MLP* is usually deleted, in order to induce the *convolutional part* to learn features which are more relevant for the given problem.

- These convolutional layers mimic in some way the first parts of the *visual cortex*. First layer takes as input the raw image, as *V1* region. From this point, each additional layer will focus on learning more complex and *abstract features*, that can be composed by the simpler ones. This hierarchical process emulates the *visual cortex* job (as long as we do not take into account the *movement* related tasks), where *V1* and *V2* focuses on the simplest features like *pattern recognition*, *colors*, *orientations* or *shapes*, while deeper regions like *V4* focuses in recognizing complex *geometries* and *forms* (we will study it deeply in section 3.1).

- The last *MLP* part focuses on the association processes such as *classification* or *recognition*. It is analogue to some regions of the *visual association cortex* such as the *PPA* or the *FFA*. It takes the *abstract features* generated by the convolutional part (*visual cortex*) and uses its knowledge for performing complex tasks (as recognizing faces like the *Fusiform Face Area*).

- As the *visual cortex* include tons short-paths between regions (for example, information of *V1* is directly given to *V4* or *V5*), later *CNN*s have implemented this behaviour by *skip connections*. With the exception of *RNN*s, these connections are always in the forward way, establishing a direct connection between previous layers and later ones. These connections have proven to improve the results obtained, suggesting one more time that, as most similar is the model to the human brain better performance it obtains.

- In cases where *RNN*s are included for detecting *movement* and *trajectories*, they are usually located just at the end of the convolutional part and just before the *MLP*. It differs from the *visual cortex* organization, where *temporal information* is mixed with *spatial information* from the very first regions (for example, both *V1* and *V5* work with *movement* and *trajectories*).

3 Characterizing convolutional layers

As each region of the *visual cortex* is characterized by the main functionalities of its neurons we can also characterize *convolutional layers* by measuring different statistics of them. For this purpose, there are some analysis that could be done. In these case we will be using the ones proposed by *NEFESI* (*Neuron Feature and Similarity Indexes*)[32]. This package, has been personally selected by its familiarity, since it was developed by myself as degree thesis. It includes the following analysis that will be useful for the current task:

Selectivity Indexes: Using the *top-100* images that maximize the activation of a neuron, it defines how selective the neuron is to different concepts like *color*, *class*, *object*, *part*... It is useful for comparing if these concepts are encoded in sections of the network that are comparable with the hierarchy of the *visual cortex* regions (figure 6).

Feature Visualization: Average those *top-100* images that maximize the neuron activation, in order to generate a visualization of the features that provokes this activation (figure 7).

Hierarchies of knowledge: By performing ablation analysis it search hierarchies of knowledge, for defining which abstract concepts are related between them (figure 9).

In following sections we will use these analysis for presenting how different characteristics are distributed through the network, in order to compare its knowledge representation with the *visual cortex* functionalities studied in section 2.2.1.

3.1 From simple to complex representations

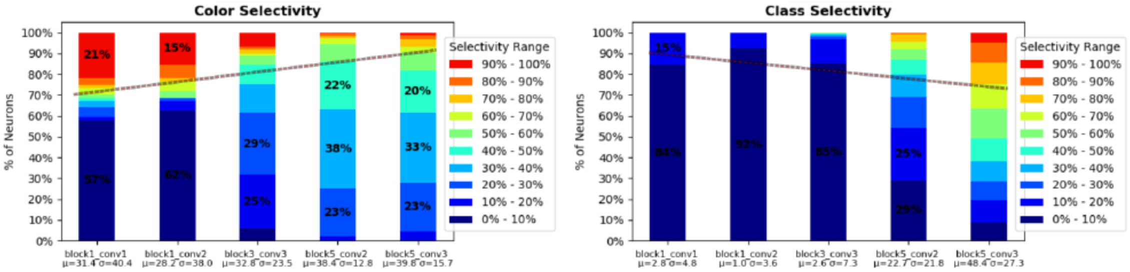

Figure 6: Concentration of neurons selective to two different concepts in each layer of *VGG16*, a simple hierarchical *CNN* with an *MLP* on the top. First layers present high activation with *color* information, while deeper layers response to more abstract concepts like *classes*.

Figure 6 shows the percentage of neurons which are highly activated by different concepts at each layer of a hierarchical *CNN* (Not using any shortcut as *skip connections*) like *VGG16*[33], trained over *ImageNet*[5]. As it shows, it is a high percentage of neurons at first layers which have high responses to *color*, like occurs in *V2* region of the *visual cortex*, and this responsiveness decreases as we go deeper through the network. Moreover, in figure 7 we can see how in these first layers, neurons that did not respond to *colors* were activated by *simple patterns* like *lines*, *orientations* or simple shapes like *circles*. It is also a very characteristic behaviour of the *V2 region* of *visual cortex*, so we could conclude that *CNN*s are able to reproduce in their first layers the most characteristic functionalities of these brain regions.

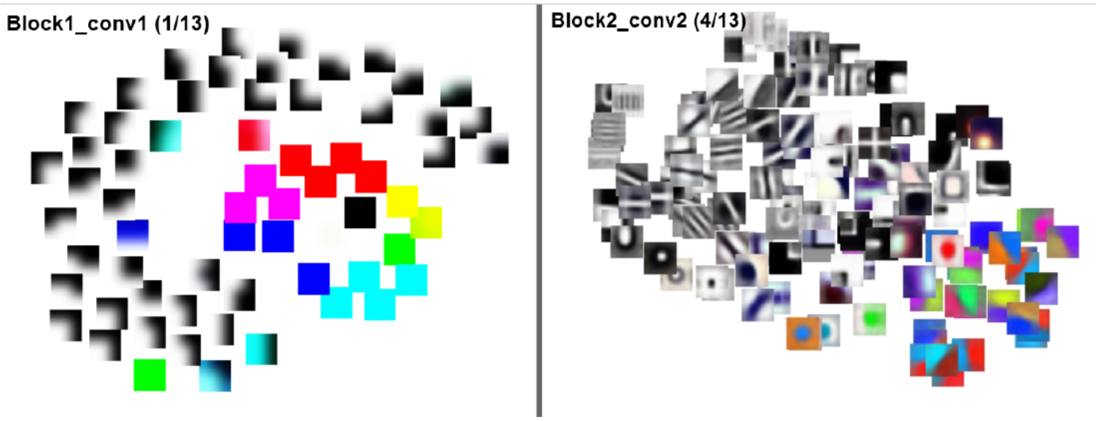

Figure 7: Average image for which each neuron maximizes its response, in function of the layer where they are placed in the first part of a *VGG16* network trained for clasifying *ImageNet*.

If we take a look on how other concepts are represented within the network (figure 6), we can observe how as deeper we go through it more increases the amount of neurons that are activated by the *abstract features* representing concrete classes (Like *birds* or *dogs*). This behaviour is quite common of some *visual associative cortex* regions like *Face Fusiform Area* (*FFA*) which present high activation for concrete *abstract concepts* like *faces* (although it can be also trained for responding to *birds*, *cars* or other *identifiable objects*). It support the hypothesis that deeper layers of *CNN*s learn knowledge representations in a way quite similar to *visual associative cortex*. In fact, they also present the same plasticity, since its neurons can be easily re-trained for changing the classes to which they respond.

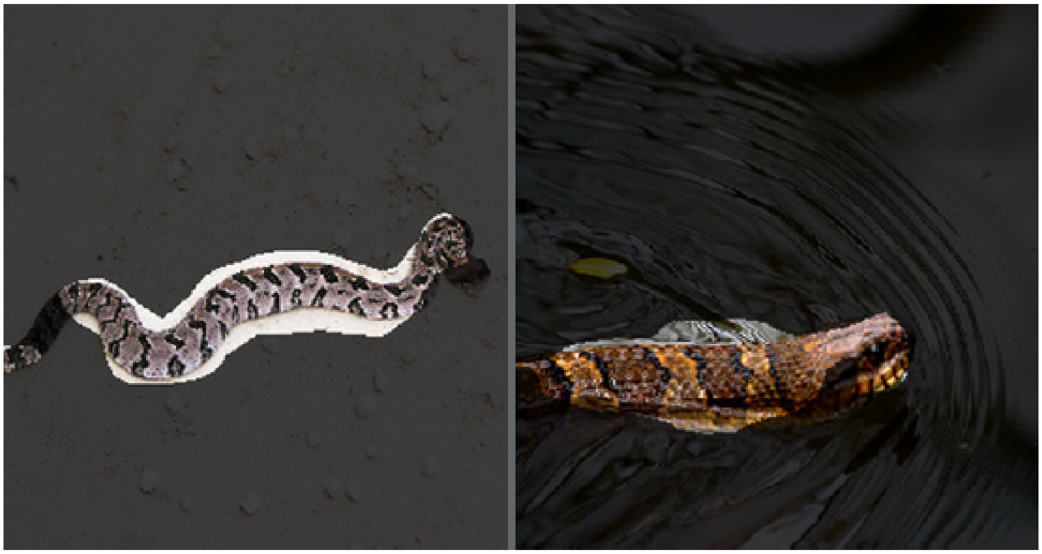

Figure 8: Parts of the image to which a singular neuron of an intermediate layer, that is selective to snakes, responds.

If we look deeper on concrete neurons (figure 8), we can see how since intermediate layers they are able to discern between *backgrounds* and the object for which they are trained. These simple segmentation ability is a function which is also present in region *V2* of the *visual cortex*, and if we consider that this neuron is detecting the flexible form of the snakes we could even relate it with the main function of the *V4* region.

It is clear then, that the knowledge representations learned by *Deep Convolutional Networks* are composed in a hierarchical way, since each additional layer is able to find more complex features than previous. These features, starts being simple like in first regions of the *visual cortex* like *V2* (*colors*, *patterns*, *orientations* or *shapes*) but, as deeper we go, they increase the complexity and the abstraction level until being able to detect complex *concepts*, like occurs in most regions of the *visual associative cortex* like *FFA*. By this reason, we can conclude that there are strong similarities between the main characteristics of *CNN*s knowledge representation and the representations that *visual cortex* and *visual associative cortex* perform. However, it is important to denote that these knowledge representations only refers to *spatial* concepts and are totally uncorrelated from any *temporal* characteristic. In the *visual cortex*, *spatial* and *temporal* characteristics are mixed since very first layers, since they detect as well *movements* and *trajectories* as *patterns* and *shapes*. Further investigation combining *CNN*s and *RNN*s would be needed for determining if these mixed models develop, in a natural way, knowledge representations which also mimic the *visual cortex* in its *temporal* characteristics.

3.1.1 The emergence of shared knowledge

On section 2.2.1.2 we have studied how different regions of the *visual associative cortex* emerge when there are some patterns that are constantly related and repeated. For example, when we are constantly visiting places and walking through them, the *PPA* emerge for favouring the *orientation* in these related situations. In the same way, it occurs with *FFA* when we are constantly matching some forms with identities, like faces with persons. This process provokes that similar tasks will be located in same regions. This special behaviour has been described by some authors as *Population Code*[34]. *Population Code* is a concept which describes how a single division of the brain is relevant for characterizing multiple concepts (As for example *faces* and *birds* in some persons *FFA*s).

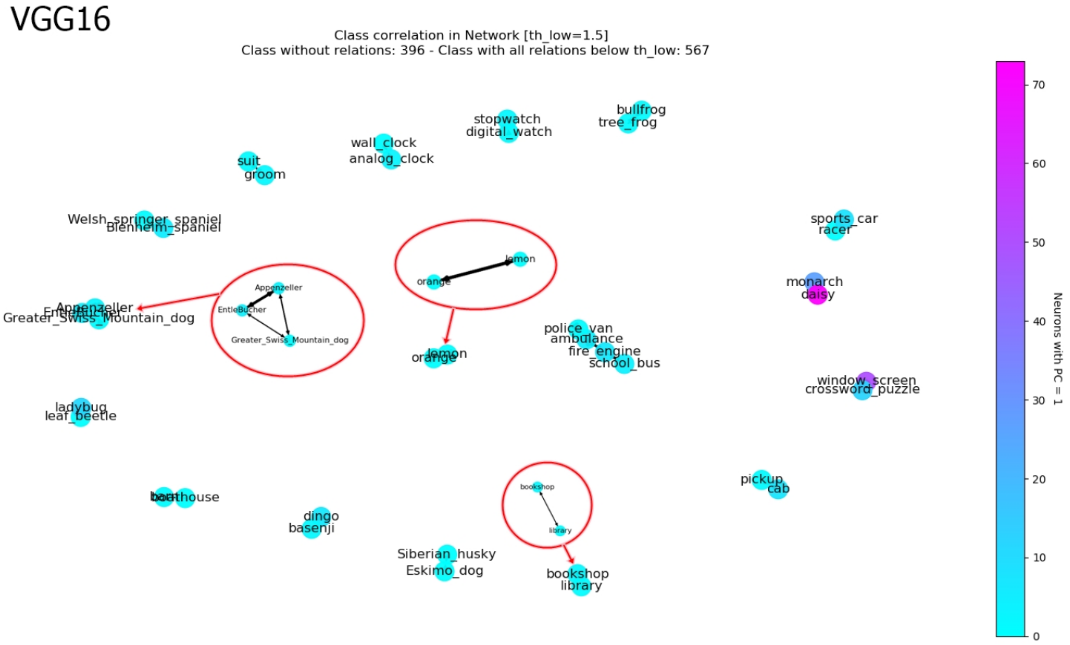

Figure 9: *Population Codes* more frequents among the neurons of *VGG16*. Two connected nodes represents two concepts that are mostly detected by the same neurons. Node color, encodes which concepts have also a high amount of neurons completely dedicated to them (*Population Code = 1*).

Figure 9 shows an study about how different concepts are also shared by same neurons in a *Convolutional Neural Network Classifier* like *VGG16*. If we look in the figure, we can observe how related concepts like *Oranges* and *Lemons* provoke high activation over the same neurons. It also occurs between different *dog breeds* or even with related vehicles, like *Police Van*, *Ambulance*, *Fire Engine* and *School Bus*.

As seen, as occurs inside *visual associative cortex* regions, in *Deep Convolutional Networks* we also have neurons with a high response to multiple related classes (Population Code > 1). It reinforces the hypothesis that, the way how knowledge emerges in *Deep Convolutional Networks* is highly related with how it emerges in *visual associative cortex*, since they both ends up with very similar representations. However, it is important to denote that *CNN*s knowledge is restricted by its unique objectives (*classification*, *face detection...*), while human brain has a myriad of parallel objectives that even change rapidly in time. By this reason, the complexity of the relations that *CNN*s can create will be always subject to that narrow objectives. For example, if a concrete classifier has learned to classify places, it will never be able to generate knowledge about indirect relations that emerge from navigating through them (as occurs in *PPA* region of the *visual associative cortex*).

4 Conclusions

In this work we have seen which are the main relations between the *human visual system* and most used *Deep Learning* models of vision. For this purpose, we have divided the *human visual system* into two main tasks: The *acquisition of images* and *the extraction of knowledge*. For the first part, we have compared how the *natural solution*, the *Eye*, is practically identical to the *artificial solution*, the *camera*. They perform the same processes in light treatment and, although they present minor differences in the way how this light is translated to electric pulses, the output obtained is quite similar. For the second and most complex part, the knowledge extraction, we have find two sub-tasks: *the abstract feature* extraction and the *association*. In both, the *human visual system* and the *Deep Convolutional Networks*, these sub-tasks do not present a clear division. The *feature extraction* is performed always by the *convolutional* part in *CNN*s, as well as by the *visual cortex* in the human brain. However, the location where the *association* task is performed is more fuzzy. Newer *CNN*s do not implement an *MLP* on the top, demonstrating that the *association* task can be performed by the *feature extractor* when these feature are too abstract. By its part, the *visual associative cortex* is a too vague term and it is not clear in which degree some deep parts of the *visual cortex* are also inside it, helping to the *association* task.

Finally, we have determined which specific functionalities of the *visual cortex* and *visual associative cortex* are mimicked by *CNN*s and which ones can not be achieved. We have found how, in general terms, the feature extraction is performed in a very similar hierarchical way, detecting simple features like *orientations*, *colors* or *short patterns* in first regions and more abstract ones like *complex forms* and *concepts* in deeper regions. We have also seen how, in the same way that related concept share same regions in the *visual associative cortex*, they also tends to share same neurons in *CNN*s. Despite that, *visual cortex* mixes *temporal* and *spatial information* in same regions, detecting *colors* and *forms* in parallel with *movement* and *trajectories*. By this reason, comparisons between both systems would never be fair, unless *Recurrent Neural Networks* schemes would be also included in the model. Further research would be needed for determining in which degree a combined *CNN-RNN* model could mimic better the mixed process that *human visual system* performs.

References

[1] Cornish-Bowden, A. & Luz Cárdenas, M. Life before LUCA. Journal of Theoretical Biology Volume 434, 7 December 2017, Pages 68-74. 2017.

[2] Vallerga S. The Phylogenetic Evolution of the Visual System. Human and Machine Vision. Springer, Boston, MA. 1994.

[3] Papert, S. The summer vision project. 1966.

[4] Krizhevsky, A., Sutskever, I. & Geoffrey, H. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 - 1097–1105. 2012.

[5] Russakovsky, O. Deng, J. Su, H. Krause, J, Satheesh, S. Ma, S. Huang, Z. Karpathy, A. Khosla, A. Bernstein, M, Berg A. C. & Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. IJCV. 2015.

[6] Yu, Q., Yang, Y., Liu, F, Song, Y. Xiang, T. & Hospedales, T. M. Sketch-a-Net: A Deep Neural Network that Beats Humans. International Journal of Computer Vision volume 122, pages 411–425. 2017.

[7] Dodge S & Karam L. A study and comparison of human and deep learning recognition performance under visual distortions. 26th International Conference on Computer Communications and Networks. 2017.

[8] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys, Vol 51, 1 - 36. 2019.

[9] Li, Y., Ouyang, W., Zhou, B., Wang, K. & Wang, X. Scene Graph Generation from Objects, Phrases and Caption Regions. 2017.

[10] Wylie, D. R. W., Gutierrez-Ibanez, C., Pakan, J. M. P. & Iwaniuk, A. N.. The optic tectum of birds: Mapping our way to understanding visual processing. Canadian Journal of Experimental Psychology, vol 63, 328–338. 2009.

[11] Phelps, H. Optic projection, principles, installation, and use of the magic lantern, projection microscope, reflecting lantern, moving picture machine. Comstock Publishing Company. 1914.

[12] Cabrera B. Human Visual System. http://obsessive-coffee-disorder.com/human-visual-system/. 2015.

[13] Budnik, V. Centre-Periphery-Difference in Low-Level Vision And Its Interactions with Top-Down and Sensorimotor Processes. PhD Thesis, Cardiff University. 2016.

[14] Kadir, T. & Brady, M. Saliency, Scale and Image Description. International Journal of Computer Vision 45, 83–105. 2001.

[15] Qiu, F. T. & Von der Heydt, R. Figure and ground in the visual cortex: V2 combines stereoscopic cues with Gestalt rules. Neuron. Vol 47 155–66. 2005.

[16] Bussey, T J & Saksida, L. M. Memory, perception, and the ventral visual-perirhinal-hippocampal stream: thinking outside of the boxes. Hippocampus. Vol 17. 898–908. 2007.

[17] Braddick, O. J., O'Brien, J. M., Wattam-Bell, J., Atkinson, J., Hartley, T. & Robert Turner. Brain areas sensitive to coherent visual motion". Perception. Vol 30. 61–72. 2001.

[18] Moran, J. & Desimone, R. Selective Attention Gates Visual Processing in the Extrastriate Cortex. Science. Vol 229. 782–4. 1985.

[19] Movshon, J.A., Adelson, E.H., Gizzi, M.S. & Newsome, W.T. The analysis of moving visual patterns. Pattern recognition mechanisms pp. 117–151. 1985.

[20] Zekir, S. The visual association cortex. Current Opinion in Neurobiology, vol 3, 155-159. 1993.

[21] Squire, L.R & Zola-Morgan, S. The medial temporal lobe memory system. Science, vol 253, 1380–1386. 1991.

[22] Ishai A., Ungerleider L.G. & Haxby J. V. Distributed neural systems for the generation of visual images. Neuron, vol 28, 979–990. 2000.

[23] Habib M. & Sirigu A. Pure topographical disorientation: a definition and anatomical basis. Cortex, vol 23, 73–85. 1987.

[24] Kanwisher N., McDermott J. & Chun M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. Neuroscience, vol 17, 4302–4311.

[25] Xu, Y. Revisiting the Role of the Fusiform Face Area in Visual Expertise. Cerebral Cortex, vol 15, 1234–1242. 2005.

[26] Aziz-zadeh, L. Fiebach, C., Narayanan, S. & Feldman, J. Modulation of the FFA and PPA by language related to faces and places. Social neuroscience, vol 3, 229-38. 2005.

[27] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

[28] Ghiasi-Shirazi, K. Generalizing the Convolution Operator in Convolutional Neural Networks. Neural Process Lett. 2017.

[29] Chung, T. & Back, A. Discrete time recurrent neural network architectures: A unifying review. Neurocomputing, vol 15, 183-223. 1997.

[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. Attention is All You Need. Annual Conference on Neural Information Processing Systems. 2017.

[31] Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A. & Tran, D. Image Transformer. International Conference on Machine Learning. 2018.

[32] Cañas, E. Neuron visualization and dissection on trained CNNs. Degree Thesis. 2019.

[33] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015.

[34] Kriegeskorte, N. & Kreiman, G. Visual Population Codes, Toward a Common Multivariate Framework for Cell Recording and Functional Imaging. 2012.