Color encoding in biologically-inspired convolutional neural networks

Ivet Rafegas[a,*], Maria Vanrell[a]

[a] *Computer Vision Center*
Universitat Autònoma de Barcelona
Edifici O, Campus UAB-Bellaterra (Barcelona) Spain

Abstract

Convolutional Neural Networks have been proposed as suitable frameworks to model biological vision. Some of these artificial networks showed representational properties that rival primate performances in object recognition. In this paper we explore how color is encoded in a trained artificial network. It is performed by estimating a color selectivity index for each neuron, which allows to describe the neuron activity to a color input stimuli. The index allows to classify whether they are color selective or not and if they are single or double color. We find out that all five convolutional layers of the network have a large number of color selective neurons. Color opponency clearly emerges in the first layer, presenting 4 main axes *(Black-White, Red-Cyan, Blue-Yellow and Magenta-Green)*, but it is reduced and rotated as we go deeper into the network. In layer 2 we find a more dense hue sampling of color neurons and opponency is almost reduced to one new main axis, the *Bluish-Orangish* coinciding with the dataset bias. In layers 3, 4 and 5 color neurons are similar between them, presenting different type of neurons detecting specific colored objects (e.g. orangish faces), specific surrounds (e.g. blue sky) or specific colored or contrasted object-surround configurations (e.g. blue blob in a green surround). Overall, our work concludes that color and shape representation are successively entangled through all the

*Corresponding author
Email addresses: irafegas@cvc.uab.cat (Ivet Rafegas), Maria.Vanrell@uab.cat (Maria Vanrell)

layers of the studied network, showing up some parallelisms with the reported evidences in primate brain that can provide some inspiration about intermediate hierarchical spatio-chromatic representations.

---

1. Introduction

Several factors such as the availability of huge image datasets annotated for object recognition tasks, as well as with more flexible hardware architectures have led to that machine learning technologies applied to computer vision entering in a new era to solve any kind of vision problems by using convolutional neural networks (CNN). These artificial networks become a flexible tool to solve vision problems of diverse nature (Lecun et al. (1998); LeCun et al. (2010)) by using a hierarchical concatenation of convolutional and max-pooling layers amongst others.

In this work we hypothesize that, considering we can find some parallelisms between layers of a trained artificial network with known evidences in the human visual cortex (Kriegeskorte (2015)), we can pursue some inspiration about how color could be encoded in beyond-opponent human visual pathway by understanding how color is encoded in layers of artificial networks. Our study is focused on one specific artificial network, it was trained by Chatfield et al. (2014) on a generic object recognition task on ImageNet ILSVRC12 dataset (Russakovsky et al. (2015)). We have selected this network architecture due to its similarity with the one used by Cadieu et al. (2014). In that work, authors proved that this kind of deep architectures start to rival the representational performance of primates in object recognition tasks.

To this end, in this paper we propose a method to explore how this artificial network is encoding color information based on the estimation of color selectivity indexes over the whole neuron population of the network. Proposed method is based on two main ideas. First, to compile the set of image patches that maximally activates a neuron. Second, to estimate a color-selectivity index on each neuron based on this set. Once we measured color selectivity indexes we

can discriminate different groups of neurons, accordingly with their ability to be color selective or not, or being selective to a single-color or to a double-color pair In this paper we will use this terminology *single* and *double* to refer to color neurons that are either selective to one single color or to a pair of colors appearing in a specific shape configuration. Note that it differs from the terminology used in Shapley and Hawken (2011) where the terms single-opponent and double-opponent are used to refer to cells responding to large areas of homogeneous color, or responding to color patterns, textures, and color boundaries, respectively. Although they could broadly have some similarities, they should not be directly equated. The classification of neurons at each layer enables interesting representational properties to be extracted, such as the amount of color tuned neurons appearing in each layer, or how color and shape are entangled through network layers, or opponency properties emerging from double-color neurons. Once we obtain the map of the network color selectivity we show a clear correlation with the color distribution of the image dataset used to train the network. Reported results provide a compelling hypothesis about color representation beyond cone-opponency.

2. A trained convolutional neural network

Convolutional neural networks are artificial networks that have been proposed by several authors (Cadieu et al. (2014); Kruger et al. (2013); Kriegeskorte (2015)) as a suitable framework to model biological vision. Although they have been designed to solve engineering problems, they take inspiration from the brain and their computations could be implemented by biological neurons. They are based on hierarchical feed-forward architectures concatenating different levels of convolutional and pooling layers. Each layer operates on their inputs to produce a representation change. More technical details regarding CNNs can be found in Sec. Appendix A.

The parallelism with biological vision is derived from the fact that a CNN presents a deep hierarchy similar to the stages in ventral stream of the human visual system. Moreover, these layers are mainly based on two kinds of operations:

(a) *a bank of convolution operations followed by a non-linearity*, which allows encoding translation-invariance of features across the visual field. This results in a set of receptive-fields with increasing size as we climb into the hierarchy; and (b) *a max-pooling operation*, which is a sub-sampling step that inserts some local tolerance and also introduce scale invariance along the hierarchy. These feed-forward architectures have similar principles to those already proposed in HMAX (Serre et al. (2007a,b)).

The main advantages of these CNNs are twofold: *flexibility for easy design* different architectures allowing different kind of vision problems to be solved; and *ability to be automatically trained* in order to learn the best weights (for all the network parameters) to achieve the best performance on a specific visual task. As we have already mentioned, these two advantages emerged due to the outstanding technological achievements in three main areas: machine learning, image-specific hardware and software, and in the construction of large labeled-image datasets.

Taking advantage of the aforementioned results, several trained CNN architectures were compared in representational performance with the primate IT cortex on the visual recognition task. The work Cadieu et al. (2014), using a kernel analysis, showed that, contrary to what happened with previous artificial architectures such as HMAX (Serre et al. (2007a,b)), current deep convolutional neural networks are starting to show important representational capabilities. Although this does not prove that these computational mechanisms are similar to the primate visual system, we cannot exclude these networks as a source of representational inspiration. Cadieu *et al.* compare two main deep CNNs, one from Zeiler and Fergus (2014) and another one from Krizhevsky et al. (2012), both trained on the same ImageNet dataset (Russakovsky et al. (2015)). Taking into account their results, here we use a trained CNN with a similar architecture to those two, the VGG-M CNN that was designed and trained on the same dataset by Chatfield et al. (2014). The architecture of this network is formed by 5 convolutional layers followed by 3 fully-connected layers. It also presents 3 max-pooling layers after the first, the second and fifth convolutional layers.

More technical details of the architecture are summarized in Sec. Appendix B.

In what follows, we propose a method to analyze how color is encoded in this trained CNN. We focus on the convolutional layers (from Conv1 to Conv5), since they are the responsible for representing color and spatial information, while we are leaving the analysis on fully connected layers for further work, since they are devoted to classification task.

3. Method

A method is proposed to explore how the network represents color information across the layers of the hierarchical architecture. The method is based on the definition of a computational algorithm that estimates the color selectivity of each individual neuron. Once the map of individual selectivity indexes is built, the corresponding conclusions can be drawn about how color is encoded through layers.

As we mentioned before in Section 2, our study is performed on the VGG-M CNN architecture (see Sec. Appendix B) that was designed and trained by Chatfield et al. (2014) on ImageNet dataset (Russakovsky et al. (2015)). This trained CNN can be analyzed by collecting the set of images that maximally activates each neuron and deriving specific measurements on these image patches and on their activation values. Additionally we can use them to visualize the intrinsic spatio-chromatic properties activating the neuron.

The following subsections present details of the ImageNet dataset, including proposing a method for visualizing the activity of each neuron, and a color selectivity index to characterize the neuron activation to color stimuli. This index will provide a global classification of neurons in order to explore how color is encoded by spatial filters composing the hierarchical architecture of the network.

3.1. ImageNet Dataset

ImageNet is a large visual dataset (Russakovsky et al. (2015)), in which images are classified according to the lexical WordNet hierarchy (Miller (1995)).

Experiments performed here are done on the ILSRVC12 version of the dataset, since it was used by Chatfield et al. (2014) for training the VGG-M network. It consists of around 1.2M images labeled in 1.000 different object categories. Images are mostly given in uncalibrated RGB color and some of them are gray-level. Images can have different sizes within the dataset but they are scaled and gray-level images are channel-wise replicated to fit into the constraints of the network, which is 224×224×3 pixels.

The dataset is built from a huge variety of scenes, including objects belonging to 1000 different categories (such as dog classes, clocks, flower classes or type of buildings amongst others). This diversity of objects can also appear with large variations in size, points of view, poses, backgrounds and a large range of lighting conditions (indoor or outdoor). Sensors used in acquisition are unknown for each image, thus there is no chance to work on any RGB calibrated color space. The CNN was trained on these uncalibrated images and they are expected to be the input to the network. However, reported results in this article are performed in an opponent-like space, where color representation is decomposed in terms of intensity and chromaticity separately. Selected space is as a linear transform on the RGB color space in order not to introduce more non-linearities besides those already included by each different sensor. This RGB to OPP (opponent space) transform is given by:

$$\begin{aligned} O_1 &= (R+G+B-1.5)/1.5, \\ O_2 &= (R-G), \\ O_3 &= (R+G-2B)/2 \end{aligned} \tag{1}$$

which is based on the one proposed by Plataniotis and Venetsanopoulos (2000) but normalizing and shifting the three axes within the range $[-1,1]$. This space was conceived to achieve some physiological inspiration on uncalibrated RGB, and it has provided interesting results in computer vision[1].

[1]Regarding this opponent transform we would also like to mention that it was shown to have a large discriminant power for image segmentation using a Principal Component Analysis in the experiments reported by Ohta et al. (1980). And it has also been shown to give the

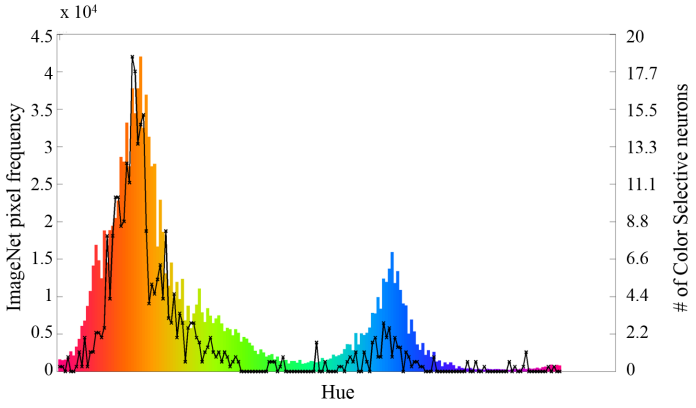


Figure 1: ImageNet hue distribution (colored bars) and Number of color selective neuron per hue (black line).

We estimated the color distribution of the complete ImageNet dataset. Colored bars plotted in Fig. 1 represent the color distribution (Y axis) based on hue-angle (X axis) computed on an O_2-O_3 plane of the opponent color space given in Equation 1. The hue-angle, h, of a given a pixel, $p = (o_1, o_2, o_3)$, is computed as $h(p) = \arctan(\frac{o_3}{o_2})$, Being $h(p) = 0$ if $p = (o_1, 1, 0) \quad \forall o_1 \in [-1, 1]$. The distribution shows a clear bias (a bimodal distribution), it peaks at orangish hues and at bluish hues. They could be due to a rich presence of brownish animals or people skin tones, and sky backgrounds, respectively. If this dataset bias correlates with natural scene statistics it will nor be analyzed. But it is normal that calibrated natural scene statistics show some kind of bias, which can vary depending on season, area or latitude (Webster et al. (2007b)).

3.2. Neuron Feature visualization

Neurons in CNNs are defined by multidimensional sets of weights (filters) which become difficult to be understood when layers are stacked as a hierarchical composition due to their high dimensionality and series of filter-composition across layers. In order to provide an image visualizing the spatio-chromatic pat-

best results in color-shape descriptors for object recognition in van de Sande et al. (2010).

tern that activates a specific neuron, we use the Neuron Feature (NF) defined in Rafegas et al. (2017). This representation visualizes neuron activity averaging the set of N-top image patches, denoted as $\{I_1, I_2, \ldots, I_N\}$, which maximally activates the neuron. The size of these patches is directly related to the corresponding receptive-field of the neuron (see Sec. Appendix B). A NF is computed as a weighted average of these image patches, where weights are proportional to the activation value produced by each image to this neuron. Thus, the NF is computed as:

$$NF(n^{L,i}) = \frac{1}{N_{max}} \sum_{j=1}^{N_{max}} w_{j,i,L} I_j \tag{2}$$

where $w_{j,i,L}$ is the relative activation of the j-th image patch, denoted as I_j, of the i-th neuron $n^{L,i}$ at layer L. The relative activation of a neuron to an image patch, $a_{j,i}$, is normalized with respect to the neuron maximum activation obtained for any patch, $w_{j,i,L} = \frac{a_{j,i}}{a_{max,i}}$ where $a_{max,i} = \max a_{k,i}, \forall k$. In this paper, we set $N_{max} = 100$ and a minimum activation value over a 70% of the maximum activation achieved by the neuron in the entire dataset. In this way the Neuron Feature allows an approximate intrinsic image to be visualized activating the neuron. In Fig.2 we can see some examples of some NFs and their corresponding set of 100 images patches used to build them. These patches are sorted decreasingly by their activation on the neuron (from left to right and from top to bottom). We can see that the appearance of the NF describes features that are mostly shared by the 100 image patches.

3.3. Color selectivity index

In this section, we introduce a computational algorithm to measure color selectivity of an artificial neuron. Color selectivity is the property of a neuron that highly activates when a particular color appears in the input image and, on the contrary, gives a low activation when this color is not present.

We propose to compute the color selectivity index of a neuron by estimating the variation between its global activation to color patches with respect to its global activation to their corresponding gray-level patches. The color selectivity index of a neuron is mathematically defined as the complement to 1 of the ratio

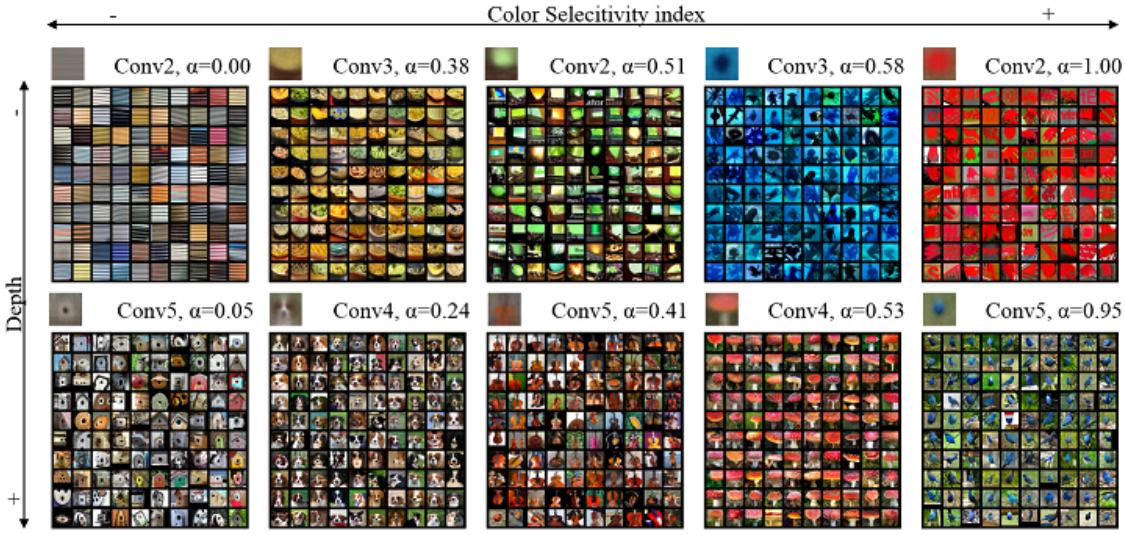


Figure 2: NFs and their 100-top image patches corresponding to 10 neurons of different layers. Neurons in shallow layers are described in two first rows, while for the ones in deeper layers, through the last two rows. The 1st and 3rd row correspond to NFs showing the intrinsic properties that may describe the neuron activity of the corresponding neuron. The 2nd and 4th rows plot the set of 100-top image patches that maximally activate the corresponding neuron. Note that these patches are sorted from left to right and from top to bottom according to activation value in a decreasing order. Color selectivity index, α is given for each neuron. Note that NFs are shown with a bigger size than each image patch in order to help in the intrinsic visualization.

between the *area under the* activation *curve* (AUC) of the gray-scale versions for the N-top image patches, and divided by the AUC of the original color patches. By *activation curve* we refer to the curve defined from the set of activations achieved by the set of N-top image patches (see Sec. Appendix C). Thus, given the set of N-top images $\{I_j\}_{j=1:N}$ of the i-th neuron at layer L, we define the color selectivity index as follows:

$$\alpha(n^{L,i}) = 1 - \frac{\sum_{j=1}^{N} w'_{j,i,L}}{\sum_{j=1}^{N} w_{j,i,L}} \tag{3}$$

where $\{w_{j,i,L}\}_{j=1:N}$ are the neuron activation values to the original N-top ranked image patches, and $\{w'_{j,i,L}\}_{j=1:N}$ are the activation values obtained by the same neurons to the gray-level versions of the N-top images.

In order to maximally preserve the shape pattern of an image, we propose to

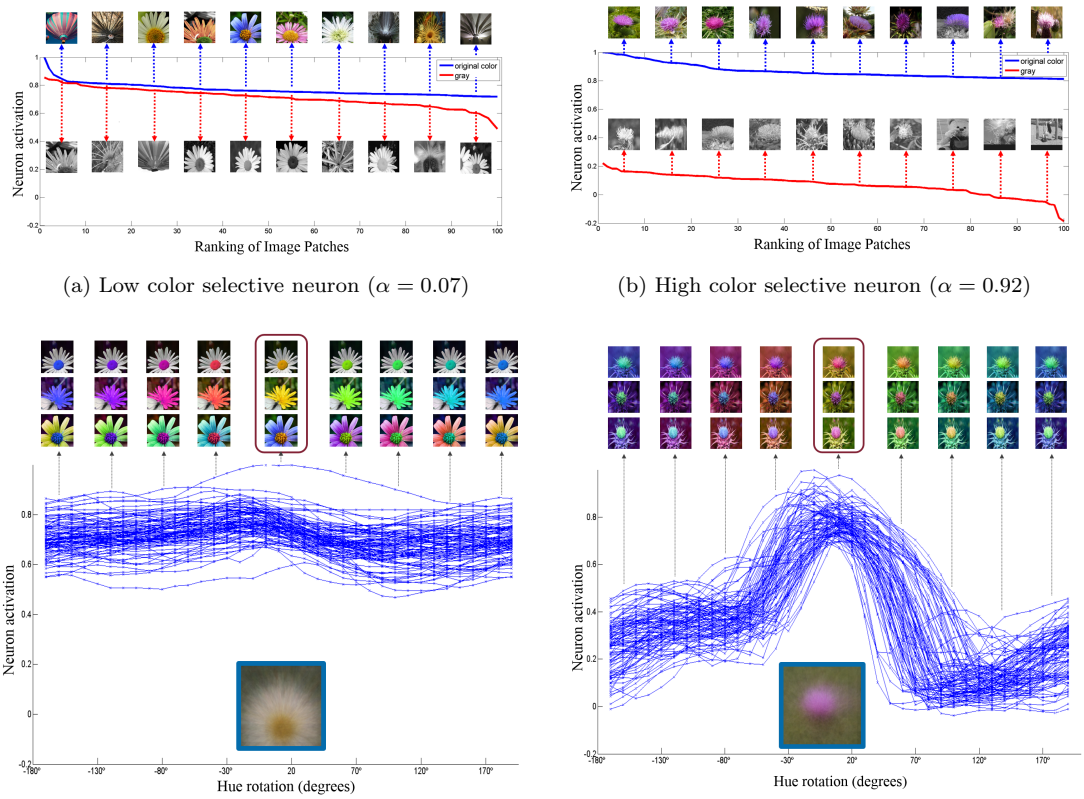
use a gray-level transformation based on the image color distribution in the OPP color space (see eq.1). We use the first eigenvector using Principal Component Analysis (PCA) on this distribution as the axis where to project each color pixel. In this way, we obtain a gray-scale image that maximizes the color image variance. We replicate this gray-level image channel-wise in order to accomplish the network constraint of receiving 3-channels. The intensity, i, of a given color pixel $p = (r, g, b)$ is computed as:

$$i = [r, g, b][e_1, e_2, e_3]^T \tag{4}$$

where e_1, e_2, e_3 are the components of the first eigenvector of the covariance matrix of the RGB pixels of the color image.

In Figs. 3a and 3b we show the neuron activity curves of two different neurons in Conv5, activation values (Y axis) are ranked in a decreasing order from left to right (X axis). We plot neuron activation curves to the N-top original image patches (in blue) and to the corresponding gray-level image versions of these N-top image patches (in red). Note that in the same X axis we are representing two different image rankings, one for color images and another for gray-level images. Neuron in 3a shows equivalent activations for both image sets (color and gray-level), while neuron in 3b plots a clear decrease in activation for gray-level images. Proposed index gives $\alpha = 0.07$ for the first neuron which is a non-color selective neuron, and gives $\alpha = 0.92$, considered color selective.

To confirm the adequacy of the proposed index, in the same figure, we explore the neuron activation to the same N-top image patches when they are transformed into different hue distributions. This transformation is achieved by a per pixel chromatic rotation on the chromaticity plane of the OPP color space, and keeping a constant intensity. In Figs. 3c and 3d we plot the neuron activation (Y axis) for each color rotation (X axis) for the same neurons shown in Figs. 3a and 3b, respectively. In these plots, we show three of the N-top image patches and some of their color transformations along the X axis. Framed with a red rectangle are the original RGB image patches. At the bottom, and



(a) Low color selective neuron ($\alpha = 0.07$)

(b) High color selective neuron ($\alpha = 0.92$)

(c) Flat neuron activation

(d) Unimodal neuron activation

Figure 3: Activation values to different image versions. (a) and (b) Ranked activation values (Y axis) to the first 100-top activated image patches (X axis), for two different neurons. Blue lines link ranked activations to the 100-top patches of a neuron. Red lines link ranked activations to the gray-level versions of the 100-top patches. (c) and (d) Activation values of the same neurons to different color transformation (rotated along hue axis) of the 100-top image patches. Blue lines are linking the activation values to all color-rotated versions of the same image patch.

framed in blue, is the corresponding NF of the neuron. Each blue line links all the neuron activation values to the corresponding images along chromaticity transformations. Again, two different behaviors emerge: a neuron that does not change its activation along the color transformations is a non-color selective neuron (flat behavior), while a neuron that clearly changes its activation in front of a color transformation is a color-selective neuron (Unimodal behavior). The mean variance computed on these activation curves for each neuron shows a

clear correlation with our proposed color selectivity index. Pearson's correlation (r) indexes through layers are 0.95, 0.85, 0.85, 0.88 and 0.88, from shallower to deeper convolutional layers.

We would like to note that this index is just measuring selectivity to color but not to what color. We work on this point in subsequent sections. The proposed method could be used to compute other selectivity indexes to focus on other properties apart from color such as spatial properties as orientation or frequency, but this is outside the scope of this paper.

3.4. Classifying neuron population

In order to analyze color coding though all network layers, we propose to use color selectivity index to classify neurons in several classes. this classification is made in three stages. The first discriminates the entire neuron population into three groups: *color selective*, *low color selective* and *non color selective*. A second stage classifies color selective neurons into two groups: *single* or *double*. And a third stage classifies double color neurons into *opponent* or *non opponent*.

The first classification is directly made by applying a threshold over the color selectivity index (α). Non color selective neurons are those with $\alpha < 0.10$, and when $\alpha > 0.25$ we label them as color selective neurons. This means that, if the AUC of a neuron activity a gray scale version of the N-top patches decreases by more than 25% with respect to their original RGB patches, it is considered a high color selective neuron. Neurons are non color selective when this AUC variation is less than 10%. Between these two groups, neurons are considered low color selective neurons. See Figs. 3a and 3b as examples of low and high selective neurons, respectively. Although the thresholds we applied can seem arbitrary, they were coherently set on the observed selectivity over the set of top ranked image patches activating the neurons, and from the behavior of the neuron activity through a chromatic transformation of the same image patch. A neuron with $\alpha > 0.25$ shows a clear unimodal behavior of its neuron activity when it is computed on the same image but presenting different hue rotations on their color pixels (see Fig. 3d), while a neuron with $\alpha < 0.10$ presents a

shows behavior (see Fig. 3c).

Within the group of color selective neurons, we distinguish two main groups: single color neurons, showing selectivity to one single color; and double color neurons, presenting with selectivity to a pair of colors. The definition of these two types of neurons was already introduced in Sec. 1. Classification is performed by fitting a Gaussian mixture model on their NF hue-angle distribution using an Expectation-Maximization (EM) algorithm. Each fitted univariate Gaussian is defined by its mean and covariance. With this fitting process, each Gaussian mean gives to which color (hue) the neuron is selective to; while its covariance indicates the amount of different hues included in a single Gaussian. The EM algorithm requires to fix the number of Gaussians beforehand. Since the higher is the number of allowed Gaussians, the better the fitting error is, we use the Elbow method to set the number of Gaussians for each neuron, it is the minimum number that achieves a mean squared error (MSE) (between the fitted and the original distribution) which differs by less than 10% of the global minimum (evaluated over 1 to 4 Gaussians). This step allows achieving one (a pair of) representative hue (hues) for each single (double) color neuron, respectively. Selectivities to three or four colors was never found in a neuron.

Once we have the color-map of all double color neurons through the network layers, we will analyze whether specific chromatic axes emerge representing spatial color opponency which is a central property in early stages of the primate visual systems (Derrington et al. (1984); Lennie and D'Zmura (1988)). At this point we want to remark here that by opponency property we mean how color pairs in double color neurons are related to each other, without trying to report it as physiological color-opponency since it is impossible in this uncalibrated space. In Fig. 6 we plot the axis related to each double color selective neuron. To evaluate the opponency property, we compute the angular distance from each hue pair, with respect to the center of the O_2-O_3 chromatic plane. The closer to 180°, the stronger the opponency property of the neuron is.

4. Results and Discussion

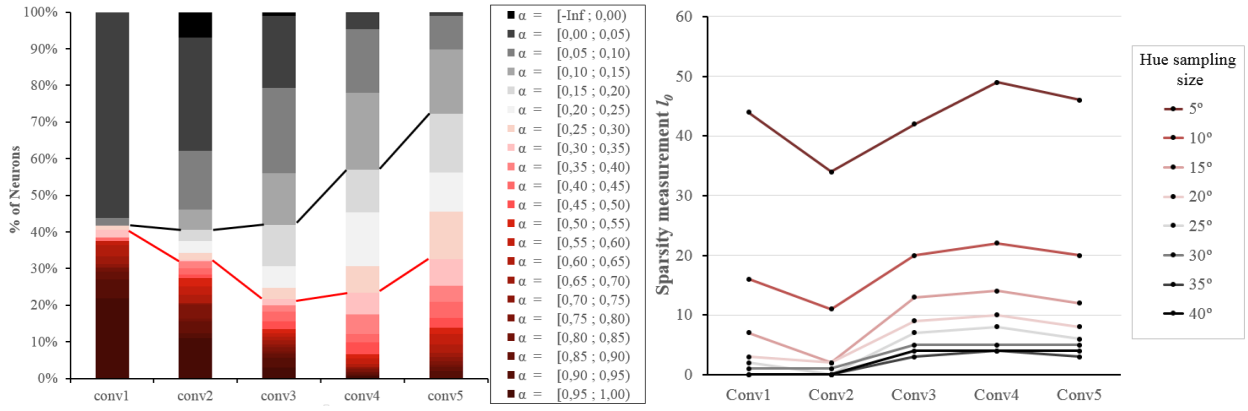
The results of computing color selectivity index on all neurons are summarized in Table 1. Neurons are classified into seven groups following the criteria defined in Section 3.4.

Selectivity #Neurons	Conv1 96	Conv2 219	Conv3 512	Conv4 512	Conv5 512
Non Color	56 (58.33%)	118 (53.88%)	225 (43.95%)	113 (22.07%)	52 (10.16%)
Low Color Sel	2 (2.08%)	28 (12.79%)	69 (33.01%)	255 (49.80%)	250 (48.83%)
Color Sel	38 (39.58%)	73 (33.33 %)	118 (23.05%)	144 (28.13%)	210 (41.02%)
Single Color	12 (12.50%)	49 (22.37%)	102 (19.92%)	134 (26.16%)	198 (38.67%)
Double Color	26 (27.08%)	24 (10.96%)	16 (3.13%)	10 (1.95%)	12 (2.34%)
Opponent	19 (19.79%)	14 (6.39%)	8 (1.56%)	1 (0.20%)	1 (0.20%)
Non opponent	7 (7.29%)	10 (4.57%)	8 (1.56%)	9 (1.76%)	11 (2.15%)

Table 1: Distribution of color and non color selective neurons through layers. Within the color selective neurons two subgroups: single color and double color, referring to the number of color the neuron is selective to. Within the double color neurons two subgroups: opponent and non opponent, depending how close are colors to present a hue-angle close to 180° or not, respectively. In parenthesis (%) percentage of neurons of the group within the layer.

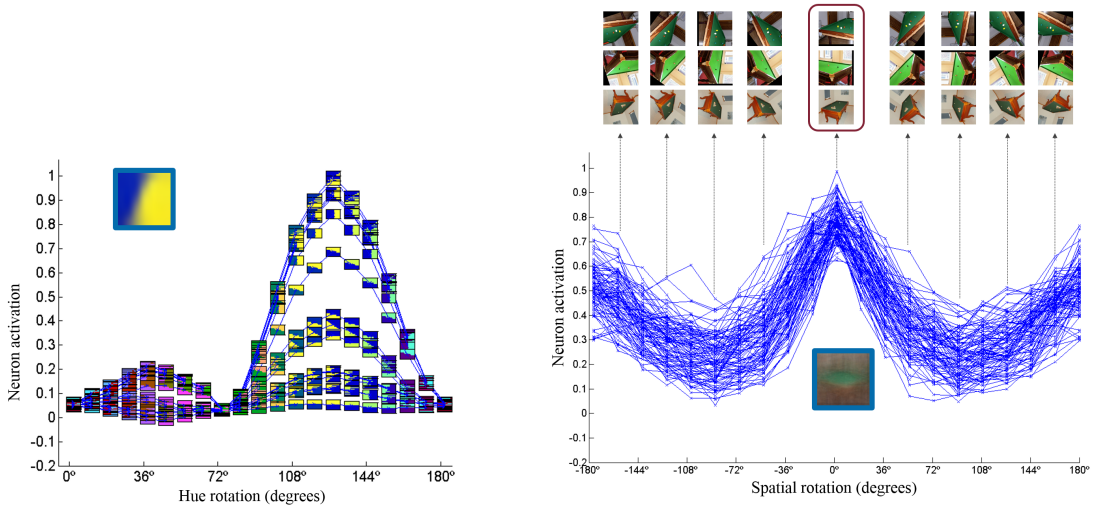
The first conclusion noted, is that there are color selective neurons in all the layers. This correlates with the idea that color is encoded all the way from V1 to IT cortex as concluded by in Shapley and Hawken (2011). A graphical representation of selectivity index values across layers and corresponding percentage of neurons is plotted in Fig. 4a. Mean color selectivity index per layer is 0.35, 0.24, 0.22, 0.24, 0.22, and 0.28 for Conv1 to Conv5, respectively, and is quite constant. But, while shallower layers have neurons with extreme color index values (either very high or very low), deeper neurons are mainly described by intermediate index values (percentage of color selective neurons in Conv4 and Conv5 surpass shallow layers).

To better understand how color and shape are entangled together at the neuron level we set up two different experiments. First, we study how the activation of a Conv1 neuron is affected by shape and color variations. The results are plotted in Fig. 5a, curves link neuron activations to images changing in color (hue rotations on the original image). Different curves have different shapes as edge orientation rotations on the synthetic original image. We can



(a) Color selectivity index distribution (b) Sparsity of hue selectivity.

Figure 4: Map of network color selectivity: how many color selective neurons and how many different colors are selective. (a) Percentage of neurons within different ranges of color selectivity indexes for each layer. Thresholds $\alpha = 0.1$ and $\alpha = 0.25$ are marked with the black and red lines, respectively. (b) Sampling of the hue space by neuron selectivity. The lower the sparsity measure is, the denser the sampling is, *i.e.*, low sparsity means a high number of different and highly distributed hues are represented by different color neurons.



(a) High color selective neuron in Conv1 ($\alpha = 0.65$) (b) High color selective neuron in Conv5 ($\alpha = 0.59$)

Figure 5: Activation variability (Y axis) of individual neurons to shape and color. (a) Each line links activations to images with egual orientation at different color variations through hue axis (X axis). Different lines for a different rotation. (b) Each line links activation to rotated versions of the same image shape. Different lines for different image patches.

see a maximal activation when both shape (orientation) and color fits with the corresponding original images (NF at top-left). Second, we perform a similar experiment in layer Conv5, but shape complexity at this level reduces the capability to analyze the extent of the conclusions. In Fig. 5b we simulate shape transformations by image patch rotations, and are applied to the 100-top image patches of the neuron (each curve represents activations to one image patch). Considering the neuron has a high color selective index ($\alpha = 0.65$), we can see the concentration of maximal activations on the original shape for all the image patches. These preliminary experiments seem to be sustaining the hypothesis of a strong entanglement through all layers, but should be performed on larger sets of neurons involving new experiments and analysis as further research.

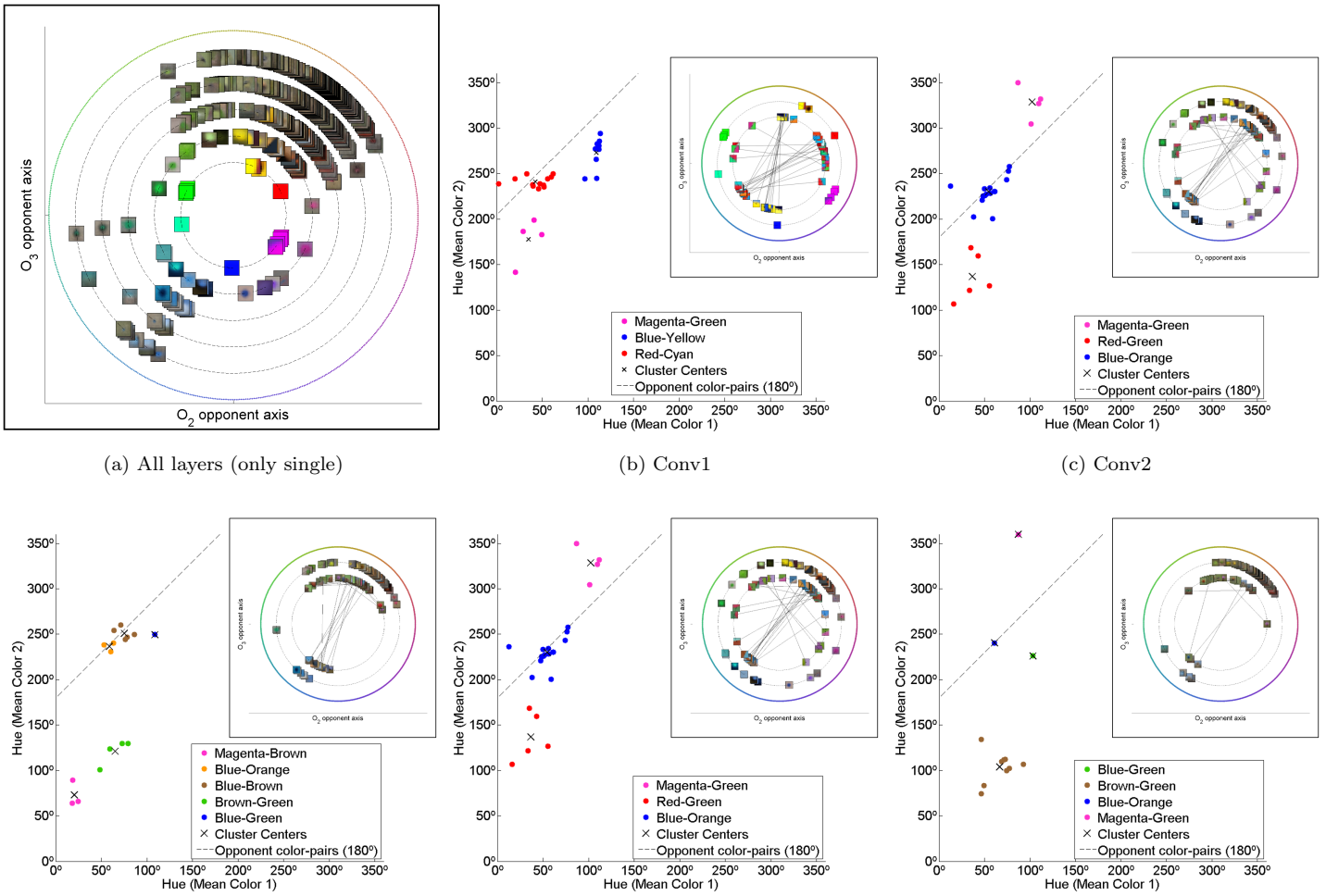
As regards single and double selective neurons, the number of single color neurons increases with depth, while the number of doubles decreases. In Fig. 6a, we plot all single color neurons along different convolutional layers, from Conv1 (inner ring) to Conv5 (outer ring). Each single color neuron is plotted at its representative hue (estimated Gaussian mean). Observe that the distribution falls in two main hue regions on deeper layers (orangish and bluish), while representatives are more distributed over hue in shallow layers. The rest of plots in the same Fig. 6b- 6f show single (outer ring) and double color neurons (inner ring) per layer. Double color neurons are plotted by their two representative hues (two estimated Gaussian means), the same NF is reproduced on both hues, and they are linked by a line to visualize their connection. Intersection of lines in similar directions outlines the emergence of color axes. A *Bluish-Orangish/Brownish* axis clearly arises in deeper layers, but this is studied in subsequent lines.

To analyze the opponency property of double color neurons, we represent their color pairs (Mean Color1 and Mean Color2) by their chromatic coordinates projected on the O2-O3 plane of the opponent space (see Fig. 6). For each of these colors we compute their hue, which is estimated as the angle between the vector formed by the projected color point with respect to the origin of the opponent space, and the O_2 axis that is taken as the origin of the hue dimension.

To distinguish the two colors of the neuron pair, Color1 denotes the one with smallest hue angle and Color2 is the other one.. In this way, perfect opponency (or 180° of angular distance between them) is represented by the location of the black dashed line (top left corner). The closer a neuron is to this dashed line, the more opponent it is. Overall, we can see from these plots that opponency decreases with depth. The maximum is found at the first convolutional layer. We performed a cluster analysis to find the main emerging axes, defined as groups of neurons sharing the same axes direction. For this purpose, we use the k-means technique and test from $K = 3$ to 7 to detect the best number of clusters using Elbow method, which is based on a ratio of the between-group variance with respect to the total variance.

In Table 2, we list the clusters obtained. We assigned an axis name to each cluster and compute its angular distance to perfect opponency. Color names used to label each axis were approximately assigned by the observation of the NFs in the cluster. From this table we can conclude two main observations: (a) in layer Conv1 all double color neurons present a remarkable opponency property (rows 1 to 3), (b) a special *Bluish-Orangish* (or a similar *Bluish-Brownish*) axis emerges from layer Conv2 up to the deepest layers (Rows 5-6). The emergence of these axes is supported by the small angular distance (in bold) to perfect opponency. Although we will refer to this axis as *Blue-Orange* (or *Blue-Brown*) the *Blue* hue presents a clear clockwise rotation with respect to *Blue-Yellow* axis found in Conv1 (see Fig. 6).

Finally, in Fig. 1 we plot the distribution of color selective neurons according to their selective hue (black line on colored bars). We combine single and double color neurons (for double color neurons we consider both hues and duplicate the single color neuron hues). We can observe a clear correlation between both distributions, presenting two significant peaks on orangish and bluish hue regions. This result confirms that color selective neurons learned by the CNN are adapted to the dataset bias, in a similar way to what happens with color bias in natural scenes that has been proved to have implications in higher color sensitivity in the human visual system (McDermott and Webster (2012)).



(a) All layers (only single)

(b) Conv1

(c) Conv2

(d) Conv3

(e) Conv4

(f) Conv5

Figure 6: Chromaticity of color selective neurons across layers. (a) Single color neurons for all layers from inner ring (layer Conv1) to outer ring (layer Conv5). In top right schemes of (b), (c), (d), (e) and (f): Single color neurons (outer ring) and double color neurons (inner ring) for layers Conv1, Conv2, Conv3, Conv4 and Conv5, respectively. Double color neurons are plotted twice (to represent double chromaticity) and linked with a line.
Graphics in left-bottom of (b)-(f) plots of emergent axes from Cluster analysis on the opponency property of double color neurons. Double neurons are represented by their pair of colors: mean hue of Color 1 in X axis (smallest hue) and mean hue of Color2 in Y axis (largest hue). Top left dashed line represents location of perfect opponent pairs (180°).

Double color neurons	Conv1	Conv2	Conv3	Conv4	Conv5
Blue-Yellow	**11.69**	-	-	-	-
Red-Cyan	**13.26**	-	-	-	-
Magenta-Green	**20.05**	32.64	89.86	101.66	65.72
Red-Green	-	56.32	-	-	-
Blue-Orange	-	**3.27**	**1.46**	-	-
Blue-Brown	-	-	**3.03**	**14.81**	**0.31**
Blue-Green	-	-	27.71	31.85	40.38
Brown-Green	-	-	87.56	107.54	100.97

Table 2: Deviation from opponency for clustered double color neurons through all layers. Neuron cluster with small deviation $< 21°$ (in bold) are proposed as opponent emergent axes.

In next sections color selectivity results are specifically analyzed in 3 groups: Conv1, Conv2 and Conv3-Conv5 together considering the similarity of their neuron population. Here we take some risk hypothesizing some parallelism of these three groups with the V1, V2 and V4/PIT/TE, suggested in the hierarchical model of color processing in macaque cerebral cortex summarized by Conway and Tsao (2009).

4.1. Layer Conv1

Trained neurons in layer Conv1 are compiled in Fig. 7a. Our classification of the neuron population in this layer can be summarized in two main groups: selective and non selective neurons around 40% and 60%, respectively. Only two neurons are found as low color selective, and as a particularity of this layer, there are more double color neurons than single. Another property that emerges by direct observation of the set of NFs is the spatial frequency shown in color and non color neurons: while color neurons only show a low spatial frequency selectivity, non color neurons present a higher diversity of different spatial frequencies. This correlates with reported evidences in human vision in Lennie et al. (1990); Schluppeck and Engel (2002).

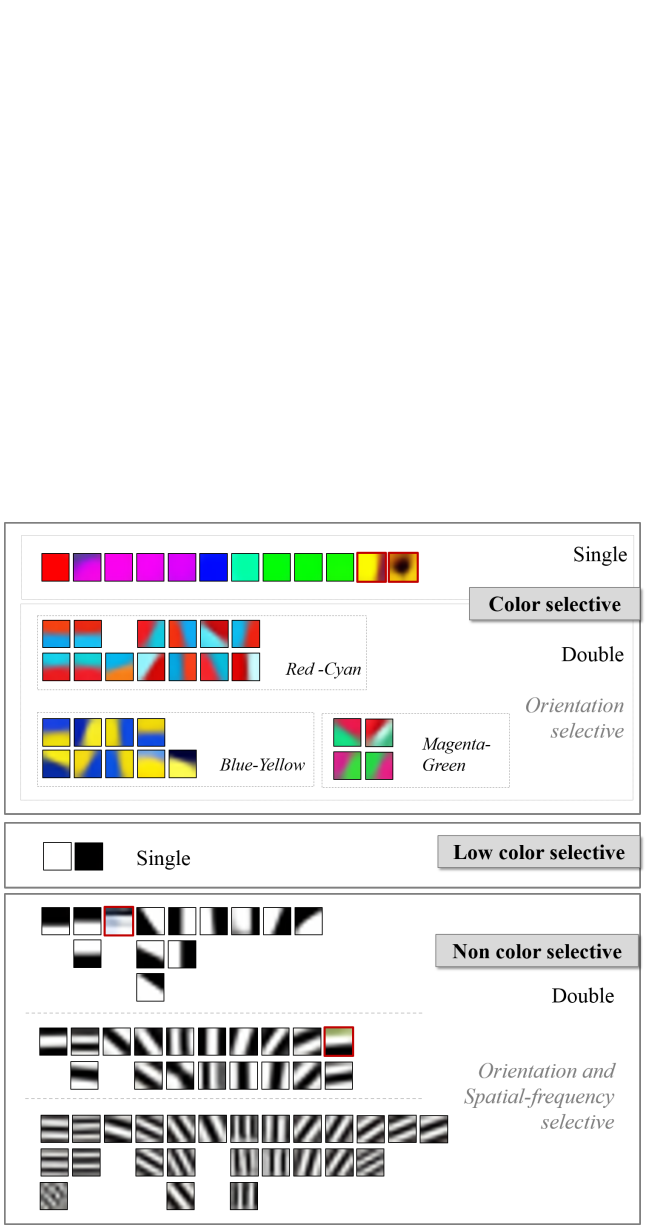
Three opponent axes emerge in this layer (see Fig. 6b). Two with a higher number of neurons *Red-Cyan* and *Blue-Yellow*, that could correlate with the findings of Derrington et al. (1984) (extensively reviewed in Lennie and D'Zmura (1988)). And a third one, *Green-Magenta*, with less neurons, but which could

be related to the fourth oponnent channel reported by Conway (2001) (*Black-White* is counted). We would like to mention here that our results are only an approximation of an uncalibrated space where labels are assigned by mere NF observation. Moreover, we would like to state that CNN training does not hold any constraint about similarities between neurons of the same layer. Then, boundaries between layers are fuzzy, and we can find neurons in second layer that could be grouped with these Conv1 neurons, as well as neurons in first layer that, duo to their spatio-chromatic properties would fit better in Conv2 (framed in red). Thus, for clarity reasons we keep our conclusions at the layer level, not trying to give a global and exhaustive classification of the color neurons.

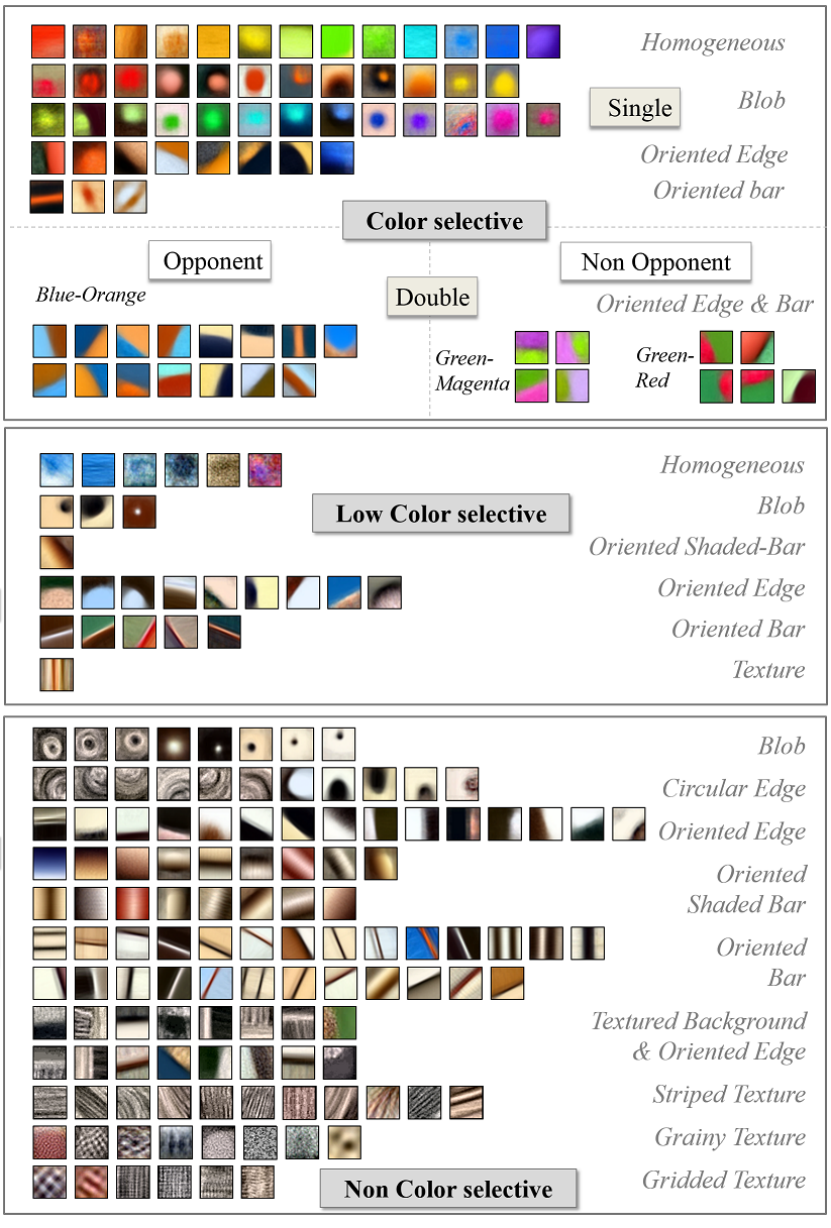
4.2. Layer Conv2

Neurons in Conv2 were classified as shown in Fig. 7b. At a first glance, we can see that non color selective neurons show an increase in shape complexity with respect to the previous layer: more complex edges as circular edges in diverse directions, oriented bars, shading effects, centered and shifted blobs or homogeneous textures and edges between textures. They are more complex features than oriented edges and basic gratings of the previous Conv1. However, they cannot be identified as object shapes like those we will find in subsequent Conv3-Conv5. This layer seems to represent surface details like different kind of textures, specific colored and oriented bars, or more complex surface boundaries (oriented c-shaped or shaded contours) in comparison to the simple oriented edges (step-like) found in Conv1.

Regarding color selective neurons, we find colored edges and homogeneous neurons as in layer Conv1. The main novelty regarding shape are colored blobs and oriented bars. Double color neurons were clustered in three main axes (see Fig. 6c): *Magenta-Green*, *Red-Green*, and *Blue-Orange*. The emergent *Blue-Orange* axis with a small deviation from opponency of 3.27° (see Table 2) is a novelty of this layer, which anticipates the edges of object shapes that will be represented in posterior layers laying on the hue bias of the dataset. Another two emergent opponent axes (*Red-Green* or *Magenta-Cyan*) seem more an extension



(a) Conv1



(b) Conv2

Figure 7: (a) 96 Neuron Features of Layer Conv 1. 38 Color selective neurons (39.58%). 12 single Color (12.50%) and 26 double opponent neurons (27.08%). 13 Red-Cyan, 10 Blue-Yellow and 3 Magenta-Green cells. (b) 219 Neuron Features of Conv2: 75 being color selective neurons (34.25%) with 54 single color neurons and 21 double color neurons (13 for the Blue-Orange, 4 for the Magenta green and other 4 for the Red-Green). Organization and label assignment in this Figure were visually performed from NF shape, with the exception of the opponent axes that come from a cluster analysis and single color neurons within the same row are sorted from the hue-angle we found using the EM algorithm for fitting their distribution by a Gaussian.

of axes in Conv1.

Another peculiarity of this layer, observed in Fig. 6a, is that color selectivity turns into a more dense sampling on the hue circle, in comparison to the rest of layers. To quantify this observation, we computed a sparsity measure l^0, studied in Hurley and Rickard (2009)[2], on the hue distribution of colors to which neurons are selective. We performed this measurement for different hue sampling sizes. The results are shown in Fig. 4b, where a clear minimum in sparsity emerges at Conv2. This continuum in hue selectivity could be related to the measurements reported by H. et al. (2008); Xiao and Felleman (2003); Webster et al. (2007a); Xiao (2014) on the existence of hue maps in V2 cortical areas reviewed in Conway (2003).

The peculiarities of the neurons of this layer can be summarized as: more complex surface features (essentially non color), a more dense hue sampling, reminiscent neurons from earlier opponent axes, and neurons defining a new *Blue-Orange* axis, which will be also found in the subsequent layers and which correlates with the dataset bias. Previous conclusions could correlate with the singular intermediate role attributed to V2 in biological systems reported in several works (Conway et al. (2010); Moutoussis and Zeki (2002); Shapley and Hawken (2011); Solomon and Lennie (2007)).

4.3. Deeper layers: Conv3, Conv4 and Conv5

The last three convolutional layers of the architecture consist of color selective neurons that seem to be highly linked to object shapes. Note that NFs present with more blurred edges of averaged images, since increase in size affects pixel-wise spatial variability.

In Fig. 8, we show an overview of all color selective neurons in these layers. Further research is required for a full understanding of color and shape representation in these layers, but we can observe 4 main groups of neurons that should be more carefully explored. First, neurons devoted to specific object

[2]Measurement l^0, defined as $l^0 = \{j : C_j = 0\}$, counts the number of zero bins in a sampled distribution, denoted as $\{c_j\}$.

shapes with a characteristic color (e.g. red and brown mushrooms, skin on faces and human bodies, and dog faces, among others). Second, neurons activated by homogeneous image areas of specific shapes simulating surround areas (e.g. sky and grass backgrounds). Third, double color neurons that can represent colored objects or object parts on specific colored backgrounds (e.g. blue bird in a green surround or ladybugs in green leafs). Fourth, single color selective neurons in which the NF does not identify a specific shape, but presenting colored regions either as a central blob or as a surround, and with strong intensity variations. From the observation of these NFs, we can also conclude that scale invariance is represented by multiples neurons representing similar shapes at different layers (different resolution), or small and large versions of the same shape within the same layer.

Double color neurons only represent 2.5% of the three layers. Some of them show clear opponency in the *Blue-Orange* or *Blue-Brown*, but some others (non-opponent) show different combinations devoted to represent green or brown surrounds jointly with different colored objects (brown, orange, blue or magenta).

Parallelism with biological systems is difficult to be established at this point, since higher-order visual areas are not as well-known as V1, and our analysis requires further research. However, later on we report some conclusions in primate visual systems that can show some similar ideas with previous conclusions, like linked color-shape or object-surround selectivity.

Multiple areas have been reported to show color selectivity (reviewed by Conway et al. (2010)). Some areas seem to combine shape and color selectivity like V4 and PIT (posterior inferior temporal cortex) while others show narrow color and saturation tuning and weak shape selectivity in TE (anterior IT). Additionally, in an earlier work, in Schein and Desimone (2011) authors stated that neurons in V4 have a high probability to be color selective to a large range of colors and white surfaces, as well as an unusual spectral property sensitive to surrounds that may play a role in figure/ground separation. In any case, a detailed study of spatial selectivity on color artificial neurons could help in shedding some more light on spatio-chromatic behaviors at these higher levels.

Most of color selective neurons of these layers are single color neurons with a huge concentration on orangish hues, as can be seen in Figs. 6d, 6e and 6f for Conv3, Conv4, and Conv5, respectively. This is due to two factors: (a) their neuron activity is related to encode objects surrounded by large backgrounds (and not simple features as in Conv1 or Conv2), and (b) the bias of the dataset to this specific orange hue. Therefore, the main difference between these layers and the previous two, is that they are devoted to select entire objects surrounded by specific backgrounds.

5. Conclusions

In this work we study how color is encoded by a trained convolutional neural network. We propose a color selectivity index to characterize the neuron activation to the presence of a specific color in the input images. We use the Neuron Feature (NF) as a tool to visualize neuron activity. We computed color selectivity on all the neurons of a CNN with five convolutional layers trained for an object classification task on a large image dataset with 1.2M annotated images in 1000 different categories. This trained artificial network was shown to have representational properties almost at the level of the primate brain. After analyzing indexes and NFs across all the network neurons, we arrived at the following conclusions:

First, a large number of color selective neurons are found through all five layers of the neuronal architecture, although index is higher in shallow layers and lower in deeper layers. Color and shape are entangled together at each color selective neuron in all layers.

Second, layer Conv1 shows a strong opponent property with three axes and a clear distinction between color and non color selective neurons. These two groups also show different spatial properties: low spatial frequency selectivity in color neurons, and high spatial frequency selectivity in non color neurons. Both conclusions show a clear correlation with human vision.

Third, layer Conv2 has two main particularities: (a) emergence of a new opponent axis in the same direction of the image dataset bias; and (b) a more

dense sampling of hue of color selective neurons (suggesting some correlation with hue maps in V2). Additionally, non color selective neurons present more complex features than the oriented edges and basic gratings of the first layer. But they cannot be seen as object shapes like those in subsequent layers. This layer seems to be representing surface details beyond its boundaries.

Fourth, layers Conv3, Conv4, and Conv5 all have color selective neurons with similar properties. Neurons are selective to colors mostly within the dataset bias, that lies on the *Blue-Orange* (or *-Brown*) axis, plus some extensions towards *Green*. Regarding the spatial activation of color neurons, we identified four main groups of color selective neurons representing different types of color shape interactions: specific object shapes (such as dog-faces, mushrooms, human body), homogeneous surround areas (such as sky or green-grass), specific object-surrounds (such blue-bird in grass or ladybug on a leaf), or generic colored shaped-blobs with strong intensity contrast. Finally, to mention that scale invariance is represented by using multiple neurons selective to different scales.

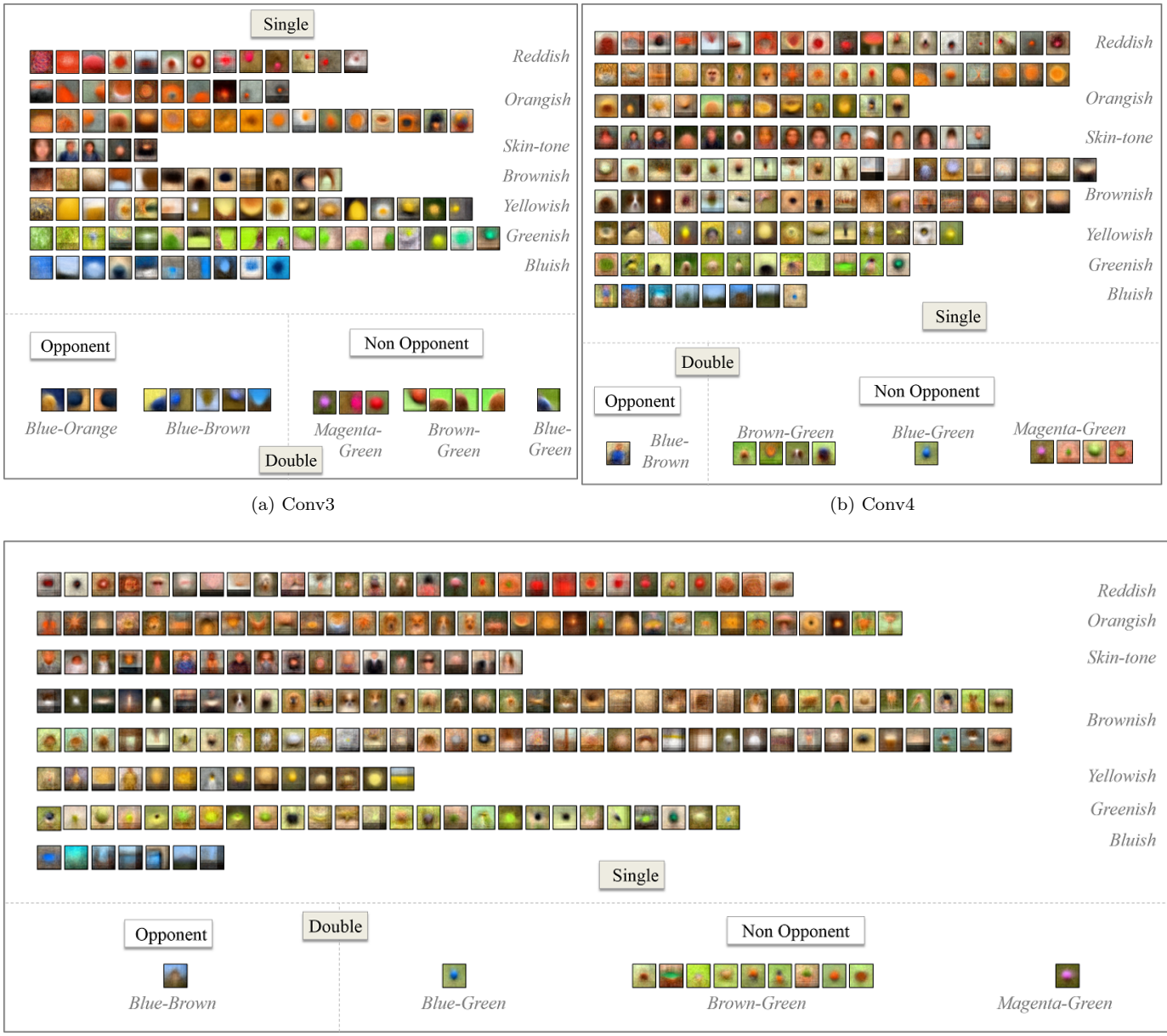
Acknowledgements. Project partially funded by MINECO Ref. TIN (TIN2014-61068-R) and Generalitat de Catalunya Ref. SGR-669.

References

Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J., 2014. Deep neural networks rival the representation of primate it cortex for core visual object recognition. PLOS Computational Biology 10, 1–18. doi:10.1371/journal.pcbi.1003963.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: Delving deep into convolutional nets, in: BMVC. doi:http://dx.doi.org/10.5244/C.28.6.

Conway, B.R., 2001. Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (v-1). Journal of Neuroscience 21, 2768–2783.



(a) Conv3

(b) Conv4

(c) Conv4

Figure 8: Neuron Features corresponding to the color selective neurons for (a) Conv3, (b) Conv4 and (c) Conv5. For each layer we differentiate between single (top) and double (bottom) neurons.

Conway, B.R., 2003. Colour vision: A clue to hue in V2. Current Biology 13, R308 – R310. doi:10.1016/S0960-9822(03)00233-1.

Conway, B.R., Chatterjee, S., Field, G.D., Horwitz, G.D., Johnson, E.N., Koida, K., , Mancuso, K., 2010. Advances in color science: from retina to behavior. The Journal of Neuroscience 30(45), 14955–14963. doi:10.1523/JNEUROSCI.4348-10.2010.

Conway, B.R., Tsao, D.Y., 2009. Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. Proc Natl Acad Sci U S A. 42, 18034—18039. doi:10.1073/pnas.0810943106.

Derrington, A.M., Krauskopf, J., Lennie, P., 1984. Chromatic mechanisms in lateral geniculate nucleus of macaque. J. Physiol. , 241–265doi:10.1113/jphysiol.1984.sp015499.

H., L., Y., W., Y., X., M., H., DJ., F., 2008. Organization of hue selectivity in macaque V2 thin stripes. Journal of Neurophysiology 102(5), 2603–2615. doi:10.1152/jn.91255.2008.

Hurley, N., Rickard, S., 2009. Comparing measures of sparsity. IEEE Trans. Inf. Theor. 55, 4723–4741. doi:10.1109/TIT.2009.2027527.

Kriegeskorte, N., 2015. Deep neural networks: A new framework for modeling biological vision and brain information processing. Annual Review of Vision Science. 1, 417–446. doi:10.1146/annurev-vision-082114-035447.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105. doi:10.1145/3065386.

Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A.J., Wiskott, L., 2013. Deep hierarchies in the primate

visual cortex: What can we learn for computer vision? IEEE Trans. Pattern Anal. Mach. Intell. 35. doi:10.1109/TPAMI.2012.272.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, pp. 2278–2324. doi:10.1109/5.726791.

LeCun, Y., Kavukcuoglu, K., Farabet, C., 2010. Convolutional networks and applications in vision., in: ISCAS, IEEE. pp. 253–256. doi:10.1109/ISCAS.2010.5537907.

Lennie, P., D'Zmura, M., 1988. Mechanisms of color vision, in: CRC Crit. Rev. Neurobiol.. chapter 3, pp. 333–400.

Lennie, P., Krauskopf, J., G., S., 1990. Chromatic mechanisms in striate cortex of macaque. The Journal of Neuroscience 10, 649—-669.

McDermott, K.C., Webster, M.A., 2012. Uniform color spaces and natural image statistics. J. Opt. Soc. Am. A 29, A182–A187. doi:10.1364/JOSAA.29.00A182.

Miller, G.A., 1995. Wordnet: A lexical database for english. Commun. ACM 38, 39–41. doi:10.1145/219717.219748.

Moutoussis, K., Zeki, S., 2002. Responses of spectrally selective cells in macaque area V2 to wavelengths and colors. Journal of Neurophysiology 87, 2104–2112. doi:10.1152/jn.00248.2001.

Ohta, Y.I., Kanade, T., Sakai, T., 1980. Color information for region segmentation. Computer Graphics and Image Processing 13, 222–241. doi:10.1016/0146-664X(80)90047-7.

Plataniotis, K., Venetsanopoulos, A., 2000. Color Image Processing and Applications. Springer.

Rafegas, I., Vanrell, M., Alexandre, L.A., 2017. Understanding trained CNNs by indexing neuron selectivity. ArXiv e-prints. arXiv:1702.00382.

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., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 211–252. doi:10.1007/s11263-015-0816-y.

van de Sande, K., Gevers, T., Snoek, C., 2010. Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1582–1596. doi:10.1109/TPAMI.2009.154.

Schein, S.J., Desimone, R., 2011. Spectral properties of V4 neurons in the macaque. VR 51, 701–717.

Schluppeck, D., Engel, S.A., 2002. Color opponent neurons in V1: A review and model reconciling results from imaging and single-unit recording. Journal of Vision 2, 5. doi:10.1167/2.6.5.

Serre, T., Oliva, A., Poggio, T., 2007a. A feedforward architecture accounts for rapid categorization. PNAS Proceedings of the National Academy of Sciences 104, 6424–6429. doi:10.1073/pnas.0700622104.

Serre, T., Wolf, L., Bileschi, S.M., Riesenhuber, M., Poggio, T.A., 2007b. Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Analysis and Machine Intelligence 29, 411–426. doi:10.1109/TPAMI.2007.56.

Shapley, R., Hawken, M., 2011. Color in the cortex: Single- and double-opponent cells. VR 51, 701–717. doi:10.1016/j.visres.2011.02.012.

Solomon, S., Lennie, P., 2007. The machinery of colour vision. Nature Review Neuroscience 8(4), 276–286.

Webster, M.A., Mizokami, Y., Webster, S.M., 2007a. Hue maps in primate striate cortex. NeuroImage 35, 771–786. doi:10.1016/j.neuroimage.2006.11.059.

Webster, M.A., Mizokami, Y., Webster, S.M., 2007b. Seasonal variations in the color statistics of natural images. Network: Computation in Neural Systems 18 (3), 213–233. doi:10.1080/09548980701654405.

Xiao, Y., 2014. Hierarchy of hue maps in the primate visual cortex. Journal of Ophtalmic & Vision Research 1, 144–147.

Xiao, Y., Felleman, Y.W.D., 2003. A spatially organized representation of colour in macaque cortical area V2. Nature 421, 535–539. doi:10.1038/nature01372.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: ECCV. doi:10.1007/978-3-319-10590-1_53.