UNDERSTANDING TRAINED CNNS BY INDEXING NEURON SELECTIVITY

Ivet Rafegas & Maria Vanrell
Computer Vision Center
Universitat Autònoma de Barcelona
Bellaterra, Barcelona (Spain)
{ivet.rafegas, maria.vanrell}@uab.cat

Luís A. Alexandre
Department of Computer Science
Universidade da Beira Interior
Covilhã , Portugal
lfbaa@ubi.pt

ABSTRACT

The impressive performance and plasticity of convolutional neural networks to solve different vision problems are shadowed by their black-box nature and its consequent lack of full understanding. To reduce this gap we propose to describe the activity of individual neurons by quantifying their inherent selectivity to specific properties. Our approach is based on the definition of feature selectivity indexes that allow the ranking of neurons according to specific properties. Here we report the results of exploring selectivity indexes for: (a) an image feature (color); and (b) an image label (class membership). Our contribution is a framework to seek or classify neurons by indexing on these selectivity properties. It helps to find color selective neurons, such as a *red-mushroom neuron* in layer conv4 or class selective neurons such as *dog-face neurons* in layer conv5, and establishes a methodology to derive other selectivity properties. Indexing on neuron selectivity can statistically draw how features and classes are represented through layers at a moment when the size of trained nets is growing and automatic tools to index can be helpful.

1 INTRODUCTION

In parallel with the success of CNNs to solve vision problems, there is a growing interest in developing methodologies to understand and visualize the internal representations of these networks. How the responses of a trained CNN encode the visual information is a fundamental question for computer and eventually for human vision.

Several works have proposed different methodologies to address the understanding problem. Recently, in Li et al. (2016) two main groups of works are mentioned. On one side those works that deal with the problem from a *theoretical* point of view. These are works such as Montavon et al. (2011) where kernel sequences are used to conclude that deep networks create increasingly better representations as the number of layer increases, Paul & Venkatasubramanian (2014) which explains why a deep learning network learns simple features first and that the representation complexity increases as the layers get deeper, Goodfellow et al. (2014) where an explanation for why an adversarial example created for one network is still valid in many others and they usually assign it the same (wrong) class, or Arora et al. (2014) that presents algorithms for training certain deep generative models with provable polynomial running time. On the other side, an *empirical* point of view, which comprises approaches that pursuit methodologies to visualize intermediate features in the image space, or approaches that analyze the effect of modifying a given feature map in a neuron activation. Our work is framed in the first subset of empirical approaches.

Visualizing intermediate features seeks to describe the activity of individual neurons. This description is the basis of this work hypothesis that is based on the idea that a proper understanding of the activity of the individual neurons allow us to draw a map of the CNN behavior. This behavior can be understood either in terms of relevant image features or in terms of the discriminative power of the neurons across the full architecture.

The first and most obvious way to describe the activity of a single neuron is given by the inherent set of weights of the learned filters. These weights can be used to compare neurons between them, either within the same layer or versus neurons in similar CNNs which have been trained under different initialization conditions, as it is proposed by Li et al. (2016). A direct visualization of these weights is intuitive when they belong to neurons of a first convolutional layer. However, when layers are stacked, that intuition disappears and the capability to understand the neuron activity is lost.

A second method to describe neuron activity is projecting the filter weights into the image space, trying to get the inherent feature that maximally activates the filter. The projection can be computed by composing the inversion of the layer operators under a specific neuron towards the image space: this was called a Decoded Filter (DF) in Rafegas & Vanrell (2016). The resulting image represents an estimation of the feature that should highly activate such neuron. The disentangling algorithm that inverts the filter would give a good estimation of the feature image if most of the layer operators were invertible. However, when the number of non-invertible operators increases, the estimation becomes unintelligible. The appearance of the DFs can be seen in Fig. 1 of Rafegas & Vanrell (2016). They have also been explored by Springenberg et al. (2015) for architectures with no pooling layers since pooling is the less invertible operator. They point out the interest of obtaining such a representation, since it would allow the understanding of neuron activity independently of the input image. However, the majority of proficient CNNs contain pooling layers.

A third way to describe neuron activity is by exploring the images that maximally activate the neuron. One of the most relevant works pursuing the visualization of intermediate features, is the one proposed by Zeiler & Fergus (2014), where they project intrinsic features of neurons from the image that have provoked a maximum spike to a certain neuron, the network representation is projected into the image space by isolating them in the deconvolution approach Zeiler et al. (2010). By observing different projections that maximally activate a certain neuron they get the intuition about the main features learned on the network. Later on, in Springenberg et al. (2015) the guided backpropagation improves the deconvolution approach by a new way of inverting rectified linear (ReLu) nonlinearities, achieving better visualizations of the activations. These approaches present a main drawback, their feature visualization is image-specific, since the maximum activation of a neuron not always generalize the intrinsic feature of the neuron. To solve this problem, in some works instead of using the image that provokes the maximum activation, they use optimization techniques to generate an image that maximizes the activation. The key point of these works is using an appropriate regularization in the generation process, otherwise, the resulting image appearance is unrealistic and difficult to understand. Simonyan et al. (2014) propose a method to generate an image which is representative of a certain class by maximizing the score of this image to be classified in a certain class (or highly activates the specified neuron) with an L_2-regularization. A similar work was performed afterwards in Yosinski et al. (2015) but taking advantage of combining three different regularizations to achieve more recognizable images. Although they have explored different regularizations to achieve more realistic intrinsic feature representations, their visualizations present important artifacts that complicate the understanding of the intrinsic property.

Finally, other works focus on proposing approaches able to reconstruct the input image given a feature map, going further of analyzing the individual neuron activity. Mahendran & Vedaldi (2015) make use of optimization algorithms to search for an image whose feature map best matches a given feature map by incorporating natural image priors. Contrary, in Dosovitskiy & Brox (2015), the authors propose to reconstruct the input image from its feature maps of a given convolutional network by training a new deconvolutional network to learn filter weights that minimize the image reconstruction error when these filters are applied to the image feature maps. With this approach they are also able to get an image reconstruction with natural priors.

In the second subset of empirical approaches, Alexey Dosovitskiy (2015) train a generative deconvolutional network to create images from neuron activations. With this methodology, the variation of the activations enables the visualization of the differences in the generated images. A similar analysis is done by Aubry & Russell (2015), but instead of forward-propagate different activations to the image space and comparing them, they observe the changes on neuron activations when similar computer-generated images with different scene factors are introduced into a CNN. These works contribute in giving a deeper understanding on the internal CNN behavior. Both works conclude that there are specific neurons which are sensitive to color changes, point of views, scale or lighting configurations.

Likewise, in Zeiler & Fergus (2014) in this work we pursuit visualizing the intrinsic feature of a neuron by analyzing the images that maximally activates a specific neuron. However, to avoid the

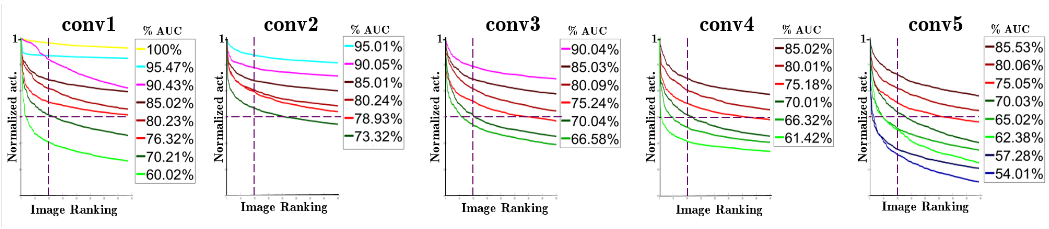

Figure 1: Normalized activations of a subset of neurons for the first 400 ranked images through all convolutional layers. For each layer we plot the normalized activation for the neurons with highest and smallest AUC (Area Under Curve), and some other examples in between these extremes. For all neurons the highest normalized activations is 1, and the percentage of AUC is computed with respect to the neuron AUC achieving the biggest area in the entire network.

lack of generality of this approach, we define the *Neuron Feature* which is not based on a single maximum activation. The Neuron Feature is a weighted average version of a set of maximum activation images that capture the essential properties shared by the most important activations and makes it not to be image-specific. Additionally, our Neuron Feature overcomes the problem of unrealistic representation we metnioned earlier, by directly averaging on the image space. In this way we achieve two main advantages: (a) keeping the properties of the natural images, and (b) providing a very straightforward approach to compute it.

Afterwards, we introduce the concept of neuron selectivity index, that is used in human vision research to characterize the response of specific cells to specific stimuli (Shapley & Hawken (2011)). This concept allows to achieve a higher level of abstraction in the understanding of a single neuron. In this work we provide two selectivity indexes which are different in their essence: a *color selectivity index* that quantifies the degree of response of a neuron to a specific color; a *class selectivity index* that quantifies the degree of response of a neuron to a specific class label. Indexes are derived from the neuron feature or directly from the set of images with maximum activations. We analyze both indexes on a VGG-M network (Chatfield et al. (2014)) trained on ImageNet (Deng et al. (2009)) and we confirm their flexibility to cluster neurons according to their index values and extract conclusions in terms of their task in the net. By selecting color selective neurons we are able to outline how color is represented by the network. Curiously we found some parallelism between color representation in the first convolutional layer and known evidences about the representation in the human visual system. Color selective-neurons also show some preferences towards specific colors which coincide with ImageNet color biases. Indexing on class selectivity neurons we found highly class selective neurons like *digital-clock* at conv2, *cardoon* at conv3 and *ladybug* at conv5, much before the fully connected layers.

2 Neuron Feature

As we mentioned in the previous section we propose to visualize the image feature that activates a neuron, whenever is possible, by directly computing a weighted average of the N-th first images that maximally activate this neuron. We will refer to it as the *Neuron Feature* (NF).

In order to build the NF we need to calculate the activations associated to each individual neuron. They need to be accordingly ranked with the rest of activations of the layer. For each neuron we select the set of images that achieve a minimum normalized activation value but constrained to a maximum number of images for practical reasons. By normalized activation we mean the value of the maximum activation of a neuron for a specific input image, which is normalized by the maximum of these values achieved by the same neuron over all the images in the dataset.

In Fig. 1 we can see the behavior of the ranked normalized responses of a subset of neurons for every convolution layer of the VGG-M CNN trained on ImageNet by Chatfield et al. (2014). The y-axis represents the normalized activation value of a single neuron to an image of the dataset. Images are ranked on the x-axis according with their activation value, from highest to lowest activation (we just plot the first 400 images for each neuron). Therefore, the first relative activation value is always 1 for all neurons and then the normalized activation values decrease monotonically. This

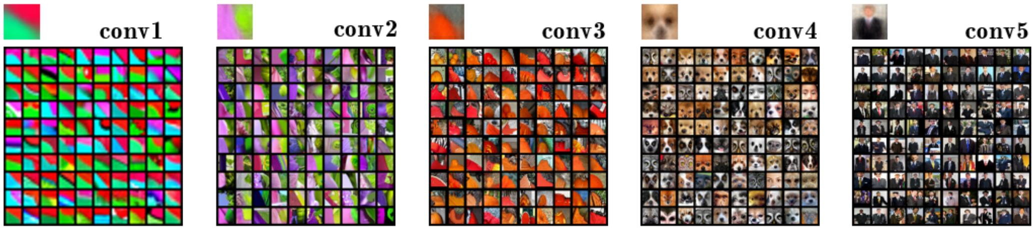

Figure 2: Neuron Feature (NF) visualizations (top) for 5 neuronsof the different convolutional layers of VGG-M with their corresponding 100 cropped images (bottom). *We scale all layers to the same size due to space constraints.*

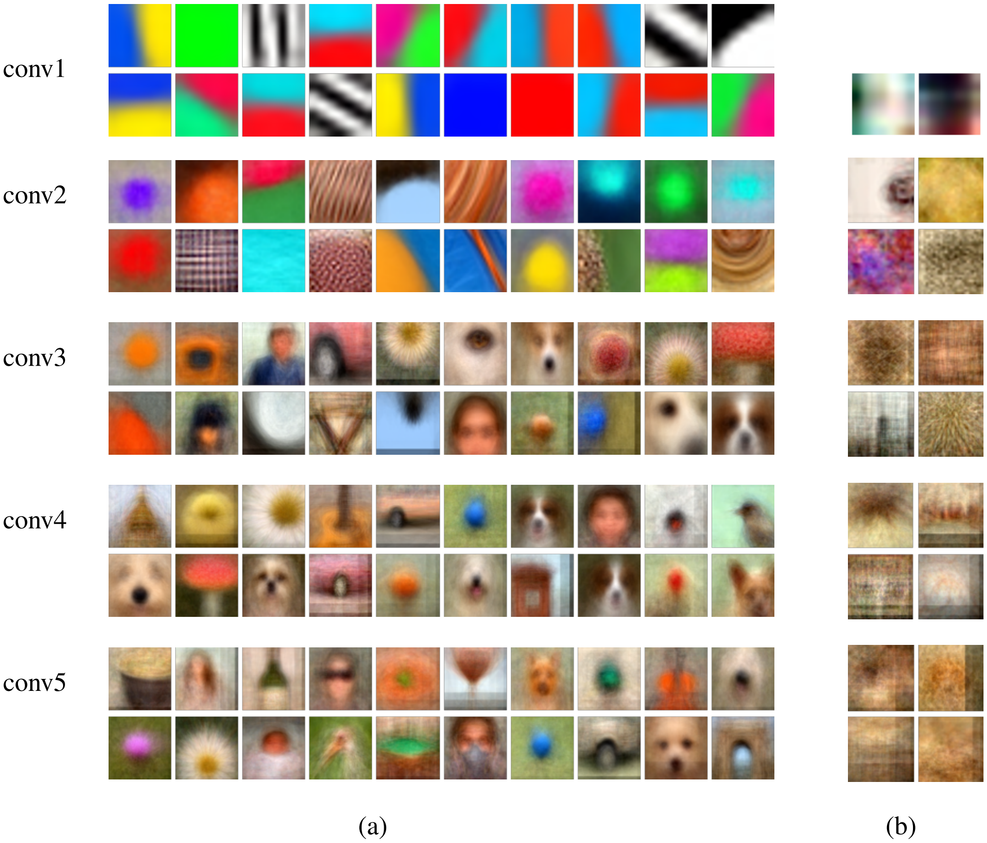

Figure 3: Examples of NFs for each convolutional layer of the network VGG-M (see section 4.1. (a) 20 examples of structured NF, (b), blurred NF. Although sizes of NF increments through layers, we scale them into the same size. Original sizes are: 7x7x3 , 27x27x3, 75x75x3, 107x107x3 and 139x139x3 for conv1, conv2, conv3, conv4 and conv5, respectively.

normalization allows to compare different neuron behaviors, from neurons which are activated by most of the images (flatter behavior), to neurons that highly activates only for a subset of images and have very little activation for the rest (steeper behavior). In this figure we also provide the percentage of area for each plotted curve. This percentage is computed over the area of the neuron that presents the maximum AUC in the entire architecture. We can observe different behaviors in all layers. In general, we can state that in deeper layers the behavior of the neurons is steeper (lower AUC), i.e. neurons highly spike for a small number of images. However, in shallower layers the behavior is flatter, i.e. neurons highly spike for a lot of images. This is an expected behavior, since the image features spiking neurons in first layers (e.g. oriented edges) are shared by almost all the images, while the features spiking shallow neurons are more selective features (e.g. faces) that only spike for specific images. The observation of the responses confirms the adequacy of our assumption to fix a minimum value for the activation and a maximum number of images to capture the most important activations for all the neurons. Similar observations have been made for other networks like VGG-S and VGG-F Chatfield et al. (2014) [1].

Thus, the NF is computed as:

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

where $w_{j,i,L}$ is the relative activation of the j-th cropped image, denoted as I_j, of the i-th neuron $n^{L,i}$ at layer L. The relative activation is the activation $a_{j,i}$ of a neuron, given a input image, with respect to its maximum activation obtained for any image, $w_{j,i,L} = \frac{a_{j,i}}{a_{max,i}}$ where $a_{max,i} = \max a_{k,i}, \forall k$.

In Fig. 2 we can see some NFs and their corresponding set of first 100 maximum activations, and in Fig. 3 (a) we can see a selected subset of 20 NF per layer. In this image we can identify specific shapes that display the intrinsic property that fires a single neuron. At first glance, we can see how in this particular network the first two layers are devoted to basic properties. Oriented edges of different frequencies and in different colors in the first layer; textures, blobs, bars and more specific curves in the second layer. The rest of the layers seem to be devoted to more complex objects. We can see that dog and human faces, cars and flowers are detected at different scales in different layers, since the size of the NF and their corresponding cropped images increase with depth. This visualization of the neuron activity can be seen as a way to visualize a trained vocabulary of the CNN that opens multiple ways to analyze the global behavior of the network from its single units. However, not all neurons present such a clear tuning to an identifiable shape. Some neurons present a blurred version of NF, such as, those in Fig. 3(b). The level of blurring is directly related to a high variability between the maximally activated images for a neuron.

At this point, we want to make a short parenthesis to relate the previous representational observations with the scientific problem about neural coding that is focus of attention in visual brain research (Kriegeskorte & Kreiman (2011)). We are referring to the hypothesis about distributed representations that encode object information in neuron population codes, that co-exist with strong evidences of neurons which are only activated by a very specific object. In line with this idea, we invite to speculate about neurons presenting a highly structured NF could be closer to localist code neurons while neurons with a blurred NF as closer to a distributed code. We return on this discussion later on at sections 4.3 and 5.

Finally, we want to add a further analysis about how neuron feature is related to the neuron activity is representing. In Fig. 4 we plot the level of the neuron responses when the input image is its own NF. We can observe a high degree of activation (in green) between the NF and the response of the net to this feature. However we have some disagreements between the NF and the neuron activations: an important example is shown in layer 2, that is curiously bigger than in layer 3 and 4. This is explained by the high number of dead neurons[2] and also by a higher presence of texture selective neurons, that is observed in Fig. 3. Another example, which is more understandable, is the clear increase of disagreement that happens through layers 3, 4 and 5, that seems to be explained by an increase in invariance that is obvious when the size of the image increases.

[1] The results are shown for a maximum number of images equal to $N_{max} = 100$ and a minimum activation value over a 70% of the maximum activation. We plot these values on Fig. 1

[2] By dead neurons we mean neurons whose activation for any input image is not enough to drive the subsequent ReLU.

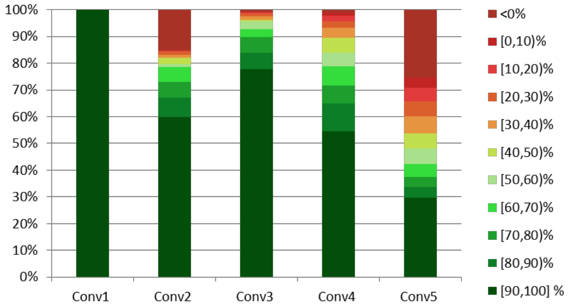

Figure 4: Number of neurons and degree of activation as a response to their own NF. Activations values are normalized to a specific range within each layer.

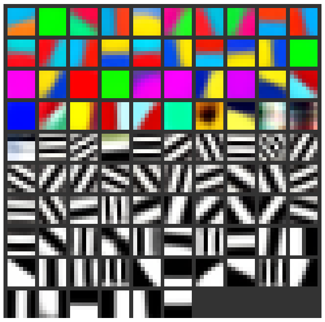

Figure 5: Conv1 NFs sorted by their color selectivity index.

3 Neuron Selectivity Index

In this section we propose to describe neurons by their inherent response to a specific property, using an index. The index has to allow to rank them in a proportional order between their response and the existence of the property in the input image. Therefore, we translate the problem of describing neuron activity to the problem of proposing methods which are able to quantify specific image facets that correlate with the degree of activation of the neuron holding such a property. A selectivity index of a single unit is a flexible an independent method for discriminating or clustering between neurons inside the same network. Selectivity indexes can be defined either for image features or for image labels. In what follows, we propose two selectivity indexes one on each group.

3.1 Color Selectivity index

Color selectivity is a property that can be proved in specific neurons of the human brain. The level of activation of the neuron when the observer is exposed to a stimulus with a strong color bias, and its corresponding low activation when the color is not present, is the object of attention in vision research that pursuits the understanding of how color is coded in the human visual system (Shapley & Hawken (2011),Conway & Tsao (2009)).

Here we propose a method to compute a color selectivity index for neurons in artificial neural networks. We propose to base it directly on the image properties of the NF we have defined above. We quantify the selectivity to a specific chromaticity directly from the color distribution of the NF. We define this index as the angle between the first principal component ($\mathbf{v}$) of the color distribution of the NF and the intensity axis ($\mathbf{b}$) of the Opponent Color Space (OPP). To compute ($\mathbf{v}$) we use a weighted Principal Component Analysis Delchambre (2014) that allows to strengthen the selectivity of small color areas. Weights are applied to each pixel in order to reinforce those pixels that are shared by most cropped images and that highly contribute to the NF. Therefore, the weights are the inverse of the standard deviation. In this way, a NF defined by cropped images with different colors will tend to be represented by a grayish image and its principal component will be close to the intensity axis in the OPP color space and it will receive a low selectivity index. We formulate this index (in degrees) as follows:

$$\alpha(n^{L,i}) = \frac{1}{90} \arccos\left(\frac{\mathbf{b} \cdot \mathbf{v}}{\|\mathbf{b}\|\|\mathbf{v}\|}\right) \tag{2}$$

Other selectivity indexes that can be derived from this, are those related to color attributes. We can easily extract color name labels using a color naming approach such as Benavente et al. (2008) and directly define color selectivity to basic names such as red, or green, among others.

3.2 CLASS SELECTIVITY INDEX

Class selectivity is a property of a neuron that can help to establish its discriminative power for one specific class or can allow to cluster neurons accordingly with the ontological properties of their class labels.

We propose a method to compute a class selectivity index for individual neurons by compiling the class labels of the images that maximally activates this neuron in a single descriptor. We define class selectivity from the set of class labels of the N images used to build the NF. To quantify this index we build the class label distribution of the full set of images. As in the color selectivity index, we weight the significance of a class label by the relative activation of its image. Thus, the relative frequency of each class c for a certain neuron is defined as:

$$f_c(n^{i,L}) = \frac{\sum_j^{N_c} w_{j,i,L}}{\sum_l^N w_{l,i,L}} \tag{3}$$

where N_c refers to the number of images, among the N cropped images activating this neuron, that belong to class c.

Given the densities for all the classes. Finally, our class selectivity index is defined as follows:

$$\gamma(n^{L,i}) = \frac{N - M}{N - 1} \tag{4}$$

where M is the minimum number of classes that covers a pre-fixed ratio, th, of the neuron activation, this can be denoted as $\sum_c^M f_c \geq th$. This threshold allow to avoid considering class labels with very small activation weight. Jointly with the index value the selectivity provides the set of M classes that describe the neuron selectivity and their corresponding relative frequency values.

Therefore, a low class selectivity index indicates a poor contribution of this neuron to a single class (minimum is 0 when $M = N$), while a high value (maximum is 1) indicates a strong contribution of this neuron to a single class. In between we can have different degrees of selectivity to different number of classes. Obviously, this index is irrelevant for the last fully connected layers in a CNN, but it allows to group related neurons across different convolutional layers.

Here we want to point out, that this index can also contribute to give some insights about the problem of how information is coded through layers, in the debate of localist and distributed neural codes we mentioned before (Kriegeskorte & Kreiman (2011)). Neurons with high class selectivity index should be in line with a localist code, while neurons with low class selectivity index should be part of a distributed code. This way the index is defined allow a large range of interpretations in between these two kinds of coding as it has been outlined in the visual coding literature.

4 RESULTS

In this section we report some empirical results to show how the proposed selectivity indexes perform and what representational conclusions we can extract from the subsets of neurons sharing indexed properties.

4.1 EXPERIMENTAL SETUP

In this paper we analyze the neurons of a CNN architecture trained on ImageNet ILSVRC dataset Deng et al. (2009) (using a subset of 1.2M images classified in 1.000 categories). We report the results for the VGG-M CNN that was trained by Chatfield et al. (2014) for a generic visual task of object recognition. The details of the CNN architecture are given in table 1. We selected this network since it has a similar structure to those which have been reported as having a representational performance that competes with human performance (as was proved in Cadieu et al. (2014)). Nevertheless, we have obtained similar results for VGG-F and VGG-S that are provided in Chatfield et al. (2014). We used the Matconvnet library provided by Vedaldi & Lenc (2015) for all the experiments.

conv1	conv2	conv3	conv4	conv5	full6	full7	full8
96x7x7 st.2, pad.0 LRN, x2 pool	256x5x5 st.2, pad.1 LRN, x2 pool	512x3x3 st.1, pad.1	512x3x3 st.1, pad.1	512x3x3 st.1, pad.1 x2 pool	4096 dropout	4096 dropout	1000 softmax

Table 1: VGG-M architecture designed by Chatfield et al. (2014), where $M \times N \times P$ corresponds to number of filters, number of rows and columns of the filters respectively. *St.* and *pad.* refers to stride and padding respectively; *LRN* is a ReLU and the corresponding pooling (*pool*) if applied.

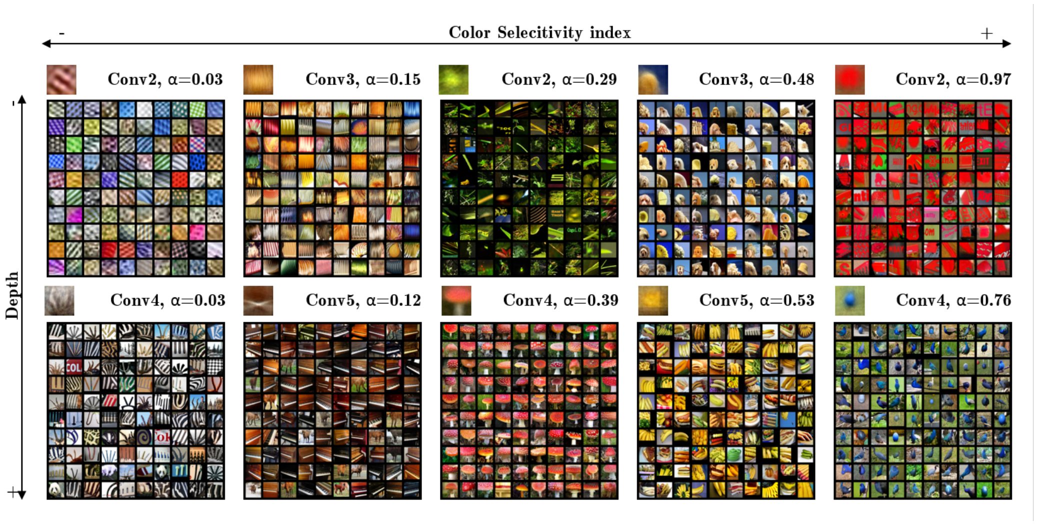

Figure 6: Neurons with different color selectivity indexes. Images in 4 rows (1st and 3rd row are NFs, 2nd and 4th rows are sets of cropped images that maximally activates the neuron).

4.2 Color selectivity

General purpose CNN architectures are usually trained on RGB color images. However there is a strong belief in the computer vision community that color is a dispensable property. The results we obtain by indexing color selective neurons make us conclude that there is no basis for such a belief. Results show that color is strongly entangled at all levels of the CNN representation. In a preliminary experiment we have tested a subset of ImageNet images with VGG-M in their original color and the same subset in a gray scale representation. Classification results show a considerable decrease: while original RGB images are classified with a 27.50% top-1 error and 10.14% top-5 error, gray scale image versions present 51.12% and 26.37% errors, top-1 and top-5 errors respectively.

In a first experiment we extract how many NFs are related to color in each convolutional layer using the proposed color selectivity index. The bars in Fig. 8 plot the relative quantity of neurons that are color selective compared to those that are not. Grey represents the ratio of neurons that do not spike for the presence of a color and reddish represent neurons that are highly activated by the presence of a color. In the graphic we can observe that shallow layers are the main responsible for the color representation on the images: 50% and 40% of neurons are color selective in layers conv1 and conv2, respectively. Nevertheless, we also still found around 25% of color selective neurons in in deeper layers. Therefore, although neurons in deeper layers tend to be color invariant, an important part of the representation is devoted to color, that reinforces the discriminative power of color in object recognition. In Fig. 6 we show some examples of NFs with different degrees of color selectivity at different layers of the network and showing the corresponding cropped images.

Regarding color representation in layer 1 we want to point out two more observations derived from the NFs (see Fig. 5): (a) selectivity to different spatial-frequencies is only tackled by gray-level neurons; and (b) four main color axis emerge (black-white, blue-yellow, orange-cyan and cyan-magenta). Curiously, these two observations correlate with evidences in the human visual system (Shapley & Hawken (2011)).

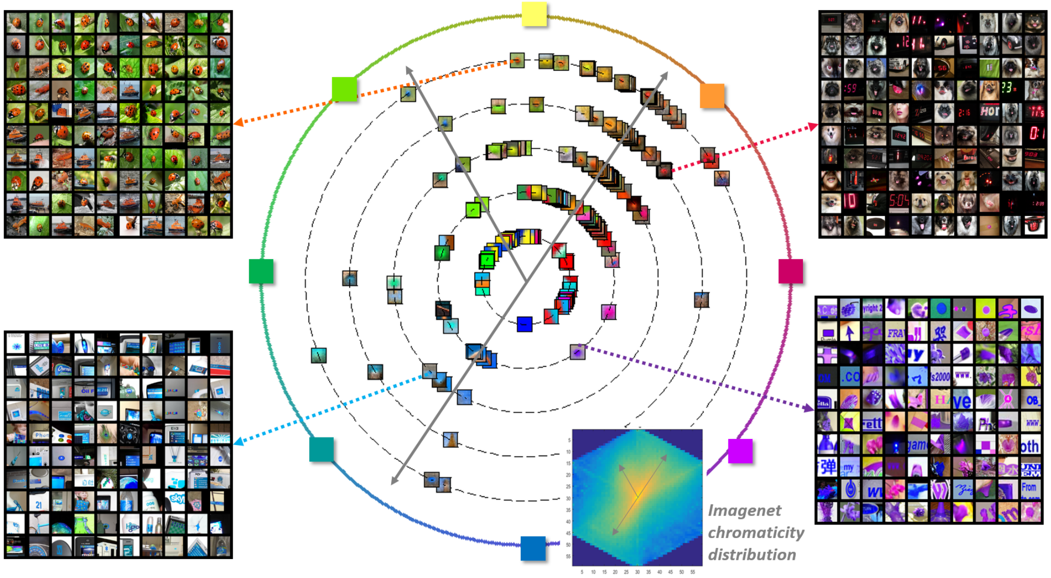

Figure 7: Distribution of color selective neurons on a hue color space through layers. Maximum activation images for 4 top color selective neurons for each layer. Dashed rings connect NFs of color selective neurons through layers, from inner ring (conv1) to outer ring (conv5).

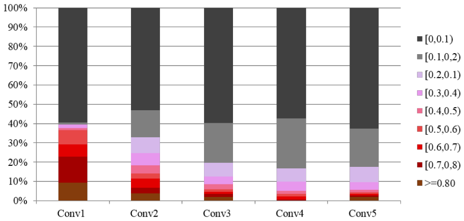

Figure 8: Number of neurons and degree of color selectivity through layers. Grayish bars are for low index values and reddish for high index values.

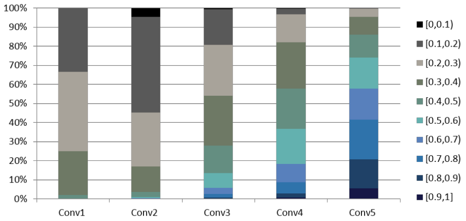

Figure 9: Number of neurons and degree of class selectivity through layers. Grayish bars are for low index values and bluish for high index values.

In a second experiment, we analyze how color selective neurons from all layers cover the color space. Figure 7 displays the distribution of color selective neurons with $\alpha \geq 0.40$. Each NF is plotted on the hue angle that represents the projection of its first principal component on the OPP chromaticity plane (red-green and blue-yellow components). Dashed rings identify different convolutional layers from conv1 (inner ring) to conv5 (outer ring) linking the NFs that belong to the same layer. We can appreciate the emergence of an axis (from orange to cyan) that connects a crowded area of color selective neurons. We can add a low population of NFs in the magenta area, that becomes more crowded on the opposite side where green and yellow selectivity has several neurons. The interest of this explanation relies on the fact that a similar distribution appears in the ImageNet color distribution, that is plotted at the bottom of the same images, where a similar interpretation in terms of emergent axes can be done. A more in depth study is required to prove this correlation, but we illustrate how neuron selectivity helps in the understanding of how a specific property is represented by the CNN.

4.3 CLASS SELECTIVITY

Following with the analysis of ranking neurons by their response to a certain property, here we focus on the proposed selectivity index that relates to image labels instead of to an image property, is the class selectivity index, which only applies for classification networks. We report the results of different experiments where we have fixed $th = 1$, which means we consider all the class labels for the $N = 100$ images that maximally activates the neuron. As we mentioned before, this index can enlighten how classes are encoded through the net layers, that again it can be related to the scientific problem of how general object recognition is encoded in the human brain. Here we hypothesize that the difference between localist or distributed codes could correlate with the idea of neurons highly selective to a single class and neurons highly selective to several classes, we resume on this later at section 5.

In a first experiment we analyze how many neurons present different degrees of class selectivity through layers. The bars in Fig. 9 plot the relative quantity of neurons that are class selective compared to those that are not. Grey represents the ratio of neurons that are not activated by a single class and bluish represent neurons that are highly activated by a single class. Opposite to what we showed about color selectivity, we found most of class selective neurons in deeper layers, and no class selectivity in shallow layers, as expected. We have moved from a very basic image property, color, to a very high level property, class label. This fact corroborates the idea that CNNs start by defining basic feature detectors that are share by most of the classes, and the neurons become more specialized when they belong to deeper layers representing larger areas in the image space and therefore more complex shapes. We start to have neurons with relevant class selectivity in layer conv3, where a 5% of neurons is quite class selective and we found some neurons with a degree of selective close to 1. These ratios progressively increase up to layer conv5 where we have more than a 50% of neurons with a class selectivity index greater than 0.6, that means that we have less than 40 different classes activating this neuron, which is a very selective ratio considering the number of classes of the ImageNet dataset. In the same layer a 20% of neurons present a high class selectivity index, than means less than 20 different classes. Further experiments should explore how this graphic evolves by moving from current class labels which are on the leaves of the ImageNet ontology towards higher nodes with more generic classes.

Secondly, we have visualized the properties of a set of images presenting different degrees of class selectivity in Fig. 10 for different levels of depth. We visualize each neuron with their NF visualization and the corresponding cropped images. We also show two *tag clouds* of each neuron. They visualize the importance of each class label. With an orange frame we plot the leave classes of the ImageNet ontology, while in the green frame we plot generic classes. This second analysis could help finding neurons that are specialized to a general semantic concept that different final classes share. Note that neurons with high class selectivity index have a set of cropped images that we can identify as belonging to the same class.

Finally we stress the utility of ranking images by selectivity indexes in Fig. 11, where we show interesting neurons in different convolutional layers that present high values for both selectivity indexes, neurons which are both, color and class selective.

5 CONCLUSIONS

In this paper we propose a framework to analyze a trained CNN by dissecting individual neurons using their indexes of selectivity to specific properties. We have proposed two properties of different nature: (a) color, that is a low-level image property that we have shown to be entangled in all the representations levels of the net; (b) class label, that is a high-level image property that can be analyzed at different levels of abstraction. We have shown that while the number of color selective neurons decreases with depth, the number of class selective neuron increases. In this line of describing the activity of individual images, we have also proposed to visualize the activity with what we have called the neuron feature (NF), that allows to arise interesting structures that are shared by the images that highly activate a neuron.

The proposed work have made us to speculate about two different ways to address the coding properties of individual neurons (localist versus distributed). Firstly, we have mentioned the possibility that a blurred NF, i.e. without a clear structure, belongs to a neuron that can be part of a distributed code

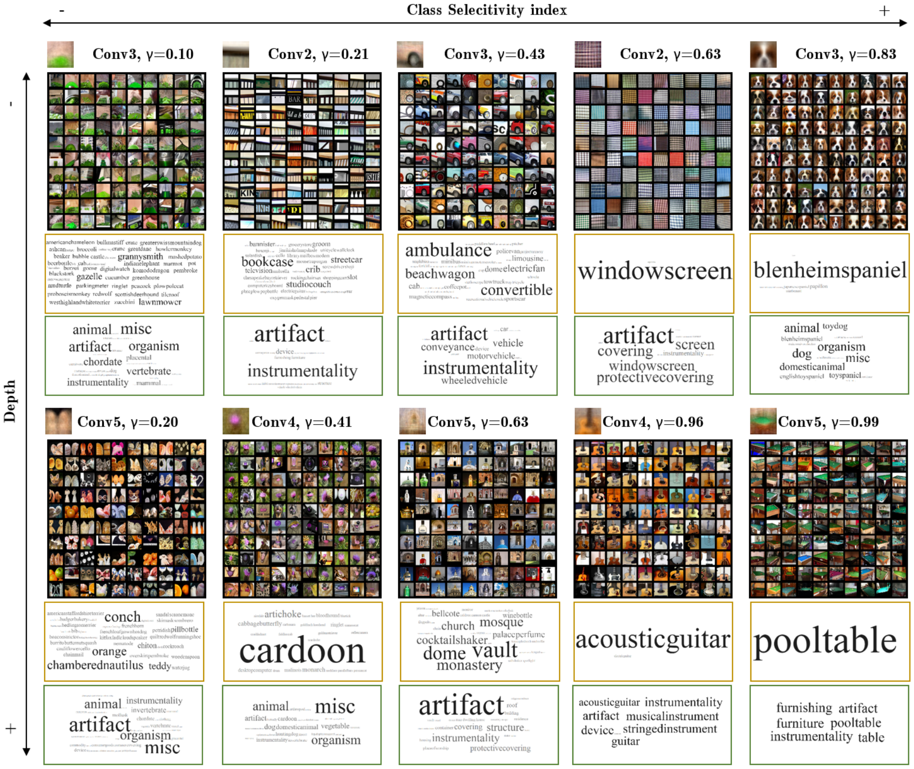

Figure 10: Neurons with different class selectivity indexes. For each neuron two images (top: NF, bottom: cropped images) and two tag clouds (top: leave classes, bottom: all classes in the ontology).

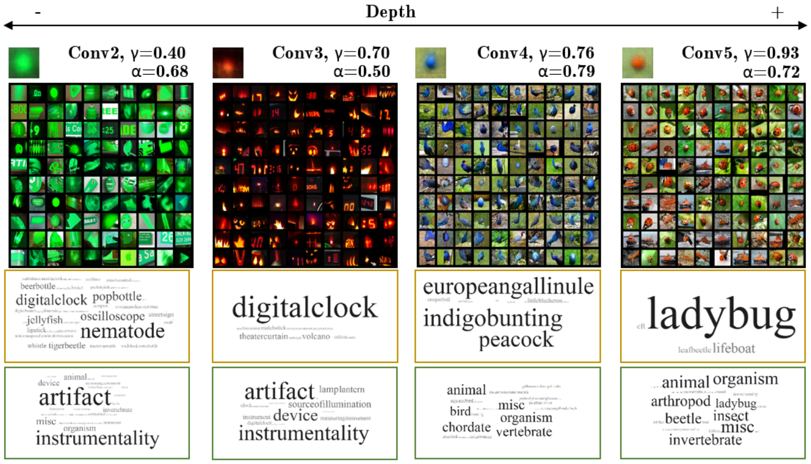

Figure 11: Examples of neurons with high color and class selectivity indexes.

where the neuron does not represent a selectivity to a single shape, maybe to diverse shapes than can be part of a code in deeper neurons. Secondly, we speculate about the possibility that neurons with high class selective index can represent a localist code, and part of a distributed when is low. In parallel, the analysis of the color selective neurons have made to arise some parallelism between color representation in the 1st convolutional layer and known evidences about the representation in the human visual system.

As further work we need to fully exploit the potential of the indexes in different CNN architectures, and defining new selectivity indexes like shape or texture, that could be a perfect complement to current ones.

ACKNOWLEDGMENTS

Project funded by MINECO Ref. TIN *(TIN2014-61068-R)*

REFERENCES

Thomas Brox Alexey Dosovitskiy, Jost Tobias Springenberg. Learning to generate chairs with convolutional neural networks. In *CVPR*, 2015.

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *ICML*, pp. 584–592, 2014.

Mathieu Aubry and Bryan C. Russell. Understanding deep features with computer-generated imagery. In *ICCV*, 2015.

Robert Benavente, Maria Vanrell, and Ramon Baldrich. Parametric fuzzy sets for automatic color naming. *JOSA*, 25(10):2582–2593, Oct 2008.

Charles F Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10, 2014 Dec 2014.

K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

Bevil R. Conway and Doris Y. Tsao. 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 (106):18034–18039, 2009.

L. Delchambre. Weighted principal component analysis: a weighted covariance eigendecomposition approach. *MNRAS*, 446:3545–3555, 2014.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. *CoRR*, abs/1506.02753, 2015. URL `http://arxiv.org/abs/1506.02753`.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL `http://arxiv.org/abs/1412.6572`.

Nicolaus Kriegeskorte and Gabriel Kreiman. *Visual Population Codes - Toward a Common Multivariate Framework for Cell Recording and Functional Imaging*. MIT Press, 2011.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John E. Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *ICLR*, 2016.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *CVPR*, 2015.

Grégoire Montavon, Mikio Braun, and Klaus-Robert Müller. Kernel analysis of deep networks. *JMLR*, 12:2563–2581, 2011.

Arnab Paul and Suresh Venkatasubramanian. Why does deep learning work? - A perspective from group theory. *CoRR*, abs/1412.6621, 2014. URL `http://arxiv.org/abs/1412.6621`.

Ivet Rafegas and Maria Vanrell. Color spaces emerging from deep convolutional networks. In *CIC*, 2016.

Robert Shapley and Michael J. Hawken. Color in the cortex: Single- and double-opponent cells. *VR*, 51(7):701–717, 4 2011.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In ICLR Workshop 2014*, 2014.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *ICLR*, 2015.

A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. 2015.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *Deep Learning Workshop, (ICML)*, 2015.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. Deconvolutional networks. In *CVPR*, 2010.