Opponent processes in representation of visual information by neurons in convolutional neural networks

Malakhova K., *Pavlov Institute of Physiology, Russian Academy of Sciences*

With remarkable similarities in the responses of neurons in the visual cortex and units of higher layers of convolutional neural networks (CNNs), CNNs are commonly considered as models of the primate visual system. However, when studied carefully, the response characteristics of artificial units diverge from what is observed in neuroscientific data. Most of the units of intermediate hidden layers of a CNN, in addition to emerging class-selectivity properties, exhibit strong negative selectivity towards some group of images from the same category. The response properties, therefore, differ from the neurons in the high-level visual cortex, where inhibition is mostly caused by images of non-relevant categories, thus pointing to differences in representation of visual information.

1. Introduction

Most widely used computer-vision models (e.g. AlexNet, VGG16, Inception-v3, ResNet, etc.), were initially trained on the ImageNet dataset [Russakovsky et al, 2015]. The dataset consists of 1000 categories, including “library”, “dishwasher”, “zebra”, “landrover”, “sea urchin”, as well as 120 categories of dog breeds to showcase fine-grained classifications.

Models trained on ImageNet dataset were shown to be useful for transfer learning [Shin et al, 2016] and to work with the novel categories without any additional training, eg. classifying paintings [Banerji & Sinha, 2016]. They also revealed remarkable similarities to the responses of neurons in the inferior temporal (IT) cortex [Cadieu et al, 2014]. Thus, representations formed in the higher-level layers of CNNs find application in various domains, including modeling visual information processing by the primate visual system.

However, this work suggests that latent representations formed in convolutional layers do not necessarily reflect the functional properties of biological neurons, such as category selectivity.

2. Methodology

Filters (or units) of a CNN are often called feature detectors based on their ability to detect the presence of a specific spatial pattern on an image. Here, the response properties of units are studied using GoogLeNet (Inception-v1) model [Szegedy et al., 2015].

Because of the prevalence of categories with dogs in the ImageNet dataset, this work studies how the concept of a dog is represented and processed in the model. For that, a combined dataset was created, which includes images of 120 dog categories of the ImageNet training data, as well as a sampling of 100 non-dog categories. The data was fed to the network and activations for every layer were collected before the application of the activation function (*ReLU*). From every activation map of a convolutional layer, in addition to the average response, minimum and maximum values were extracted (*min* and *max* pool operations).

One way to understand a neuron’s function, which is commonly used in neuroscience, is through category selectivity measure, which reflects a proportion of a unit’s category-related activity. Selectivity index was calculated on normalized activations as following:

$$Selectivity = \frac{\sum AC}{\sum AN + \epsilon},$$

where *AN* denotes normalized activation, while *AC* is a proportion of activation related to the category`s exemplars. The approach does not require class sizes to be equal. The output equals to one if all of the responses were seen in the category of interest. A filter was considered selective if its score exceeded 0.65 [Aparicio et al, 2016].

3. Results

In the intermediate layers of a network trained on ImageNet, a large proportion of filters respond to dogs’ images. When the negative activations are truncated by *ReLU* function, responses of such units will remind a tuning curve of face-selective neurons in the inferior temporal cortex [Sato et al., 2009]. However, when negative values are taken into account, it becomes possible to see that the strongest negative responses can also be elicited by images of dogs, see Fig. 1.

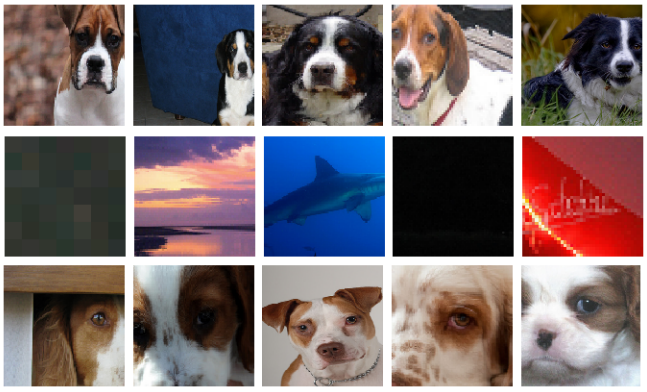

Figure 1. Examples of images that evoke a neuron of *mixed4c* layer: the strongest response (top row, μ=1226), no activation (middle row, μ=0.6), or strong negative activation (bottom row, μ=-412).

Each layer’s responses were sorted, and the content of one thousand of the strongest positive activations was compared to the content of a thousand of the strongest negative values for each filter. Besides, every filter was also assigned with a selectivity index. Figure 2 illustrates the results of the analysis for the layer *mixed4c*. If a hidden unit has properties similar to neurons in the IT cortex [Sato et al., 2009], whenever it exhibits general selectivity towards a dog category, most of the images of this category would elicit an increase in a firing rate. However, that’s not the case with artificial neurons. As we can see on Fig.2, most of the dog-selective neurons have a large number of dog images that resulted in strong negative activations.

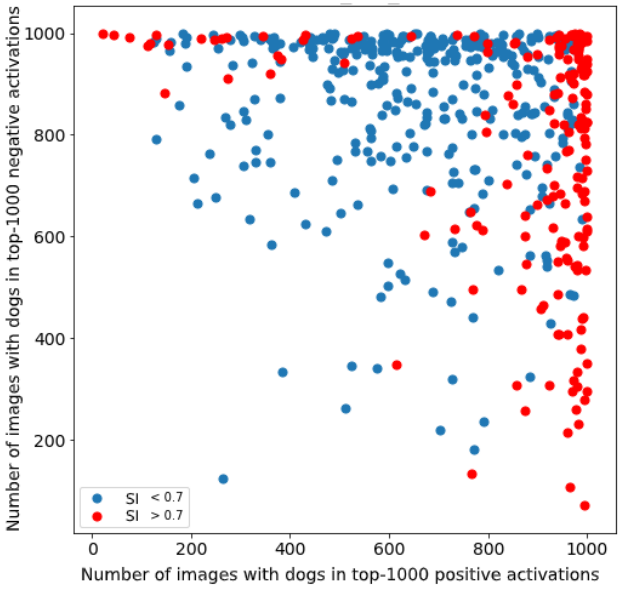

Figure 2. Presence of images with dogs in the responses of *mixed4c* layer units. Red color denotes units with a selectivity index towards dog category of 0.7 and higher.

4. Discussion

Despite the previously shown similarities between activations of neurons in the IT cortex and units in high layers of CNNs, artificial neurons encode visual information differently. In the higher-level areas of the visual cortex, neurons demonstrate sparse responses with selectivity to faces, objects, and some of the other categories.

Neurons of the CNN, in contrast, reveal dichotomy in responses. Their activations, indeed, measure the similarity of an image to a feature they detect; however, this function is not smooth. Latent space in CNNs from the early stages of the processing is influenced by the choice of a cost function and a training dataset.

The default choice of a lost function for the multiclass classification task is the cross-entropy loss function. It computes the divergence of a predicted class probability from the actual label, which is encoded as a binary vector, where one is assigned for the target class, and zeros stand for the rest of the classes. Thus, it implicitly assumes that the classes are equally spaced, meaning that a failure to distinguish a dog from a car is equivalent to confusion among dog breeds. The more confidence a node has in predicting a class, the more significantly its weights will be adjusted to avoid further mistakes.

This strategy helps to train models to differentiate categories successfully, but it does not necessarily require visually similar images to be similar in the latent space of high-level layers of neural networks. Moreover, a neuron's weights may be tuned to separate these resembling images. The most prominent example is when an individual unit, highly selective to some members of a category, is, nevertheless, inhibited by visually similar objects of the same category, and this selectivity-profile cannot be attributed to incidental differences in low-level statistics.

As a result, images from totally different categories, such as cars or cups, evoke higher activation in a dog-selective neuron than images of dogs of other breeds. These neural networks' properties may causen unstable behavior such as detecting objects in their absence or failing to recognize obvious (from a human observer's point of view) images.

Literature

1. Aparicio P.L., Issa E.B. & DiCarlo J.J. Neurophysiological organization of the middle face patch in macaque inferior temporal cortex. // Journal of Neuroscience. 2016. V. 36 № 50. P. 12729–12745.
2. Banerji, S., & Sinha, A. (2016, December). Painting classification using a pre-trained convolutional neural network. In International Conference on Computer Vision, Graphics, and Image processing (pp. 168-179). Springer, Cham.
3. Cadieu C.F., Hong H., Yamins D.L., Pinto N., Ardila D., Solomon E.A., DiCarlo J.J. Deep neural networks rival the representation of primate IT cortex for core visual object recognition // PLoS Computational Biology. 2014. V. 10. № 12. P. e1003963.
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.
5. Sato, T., Uchida, G., & Tanifuji, M. (2009). Cortical columnar organization is reconsidered in inferior temporal cortex. Cerebral cortex, 19(8), 1870-1888.
6. Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), 1285-1298.
7. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
8. Young, M. P., & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. Science, 256(5061), 1327-1331.