Representation of categories in filters of deep neural networks

Katerina Malakhova
Pavlov Institute of Physiology Russian Academy of Sciences
199034, Russia, St. Petersburg, Makarova emb., 6.
katerina.malahova@gmail.com

Abstract

Transparency in decision-making is an essential aspect of the secure and unbiased application of deep learning for classification problems. Neural networks pre-trained on one dataset can serve as feature extractors to solve various tasks. In this work, I study how categories are represented in latent space of neural networks using an example of face recognition by a network trained without an explicit category for the human person. I propose a semantic-based approach to determine if a model has pre-trained filters for a given set of classes of interest and which layer is better suited for feature extraction. The method is similar to category-selectivity measures used in neuroscience to estimate tuning curves of neurons in high-level areas of the visual cortex.

1. Introduction

Categorization of objects and events is an essential ability for decision-making and efficient operation in an environment. Convolutional neural networks (CNNs) share some functional properties with the human visual system. For example, it has been shown that features detected by the first layers of CNNs are similar to those which drive activation of neurons in the primary visual cortex [4]. However, a complete understanding of the processing that occurs in higher visual areas, as well as in deep layers of artificial networks, remains to be established. In this work, I address the problem of understanding the latent space of deep neural networks by studying category-selectivity of associated filters. I propose an approach verifying if a particular pre-trained network has weights helpful for classifying a category of interest, and if so, which layer contains the most efficient detectors. This approach can also be used for the evaluation of complexity-balance in datasets by providing a metric for comparing classes by the amount and the level of filters involved in their detection. The code of the project is publicly available on GitHub [1].

2. Interpretation of filters in deep networks

Filters of CNNs may be considered as minimal functional units as they weigh input and provide an output, which reflects how well a signal fits a particular filter. Some approaches have used visualization techniques to understand filters functional properties through selection or generation of the input that maximizes activation of the filter [2, 7, 6]. As a result, the filter can be explained qualitatively by an image, or a set of images, evoking a high response. Visualization of a preferred input as an image provides an idea of a filters function, but does not give any information about the domain to which this input may belong. However, it covers only a few picked examples, without estimating the variety of natural stimuli which activate the filter. Neurons in deep layers usually have complex functionality and interpreting it with a 2D image can be helpful, but may also be misleading.

2.1. Category selectivity measure

I propose a category-selectivity approach which uses semantic annotation to explain a filters activity. Measurements are obtained by passing a dataset containing classes of interest through the network in order to collect activations. Maximum values are computed for every activation map, and vectors of maximum-activations are then normalized for every filter:

$$AN = \frac{A - min(A)}{max(A) - min(A) + \epsilon}, \tag{1}$$

where A denotes activation tensor. Normalized activation is used to compute the metric of selectivity (equation 2), which reflect a proportion of the filters category-related activity, with the output equal to 1 if all the activation was seen in the category of interest:

$$Selectivity = \frac{\sum AC}{\sum AN + \epsilon}, \tag{2}$$

with AN denoting normalized activation and AC denoting the part of an activation related to the categorys exemplars. The approach does not require class sizes to be equal.

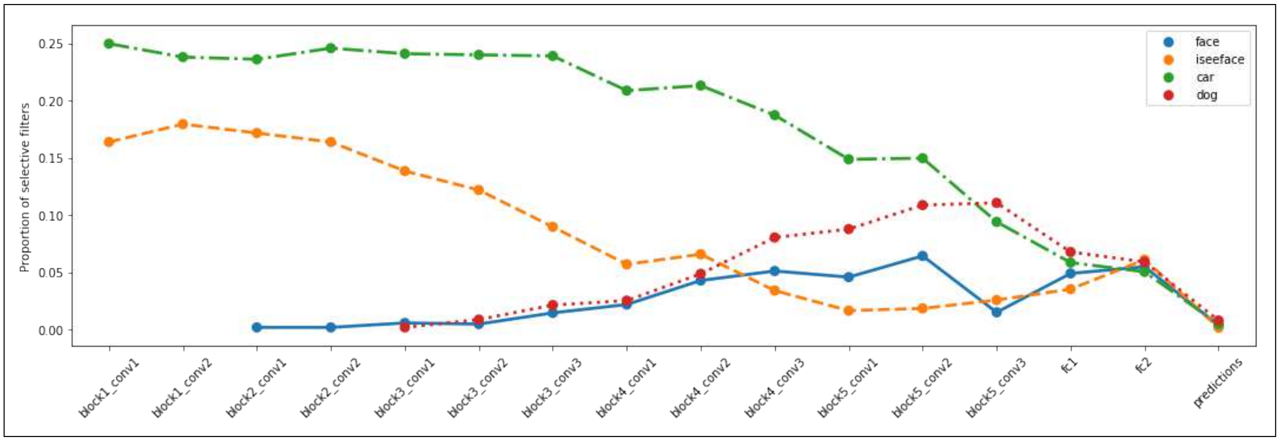

Figure 1. Proportion of category-selective filters in layers of VGG16.

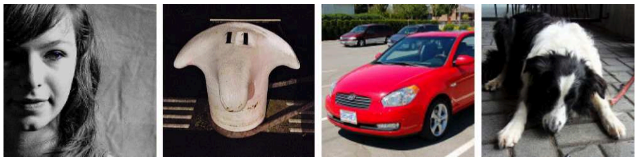

Figure 2. Example of class images (from left to right): face, i-see-face, car, dog.

2.2. Identification of face-selective filters

ImageNet (ILSVRC2012) [3] is one of the most popular datasets in machine object recognition. It includes more than 1.2 million images from 1,000 categories. Models pre-trained on ImageNet are then used for fine-tuning or feature extraction to solve custom classification tasks. Among all classes there is no category for the human person, face, or body. To understand if faces are present and processed by neural networks trained on ImageNet dataset, I created a demo dataset containing 300 random images of 4 categories (2): face, i-see-face, car, dog. Images were collected from the Internet under the relevant tags. The i-see-face category includes user images with the same name tag, referring to a pareidolia effect when face-like shapes are seen in objects, shades, clouds, etc. The choice of classes is directed here by visualization clarity and interest in distinguishing animate and inanimate objects. The dataset is tested on VGG16 [5] to illustrate the approach.

Figure 1 illustrates relative proportion of category-selective filters in each of the networks layers, with a threshold of 0.01 for selectivity index. Filters activated by cars and pareidolia-type images are vastly represented in lower layers, which may reflect distinctive statistics of representatives of these categories. Dog detectors outnumbered face-selective filters which was expected given the distribution of output classes in ImageNet. The majority of face-related units lay in fully-connected layers (not shown on the plot), but proportionally the last convolutional blocks of VGG16 (layers block4 conv3, block5 conv1, block5 conv2 and block5 conv3) contained significant amounts of category-specific detectors. The approach also allowed for a determination of representative filters for every category. Visualization of these filters using PPGN generator [2] displayed some common properties inherited by class exemplars (Fig. 3). Images may appear corrupted, but this is noted to occur because they were produced for filters in hidden layers, where neurons do not necessarily have clear and explained preferences for visual input.

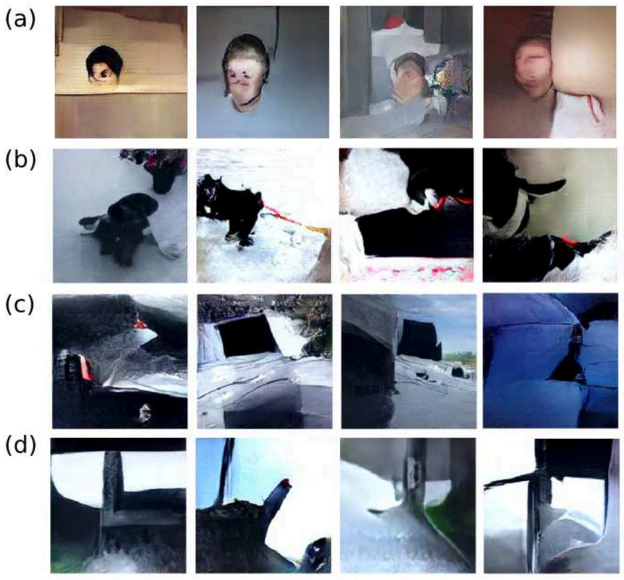

Figure 3. Example of images generated using [2] to visualize category-selective filters. (a) face category, (b) dogs, (c) cars, (d) i-see-face.

3. Discussion

Visualization of network properties through category-selectivity of individual filters helps to build a deeper understanding of functional properties of neural networks. The approach can be used to identify a particular layer which contains sufficient amounts of detectors sensitive to the category of interest. A combination of outputs from different layers that contain selective filters can then be used to achieve better performance in a classification task. This approach can also be used in psychophysiological studies to ensure that images from different categories presented in a study have similar visual complexity. This way the visual complexity can be measured as the number of filters highly responsive to a class at a given layer. Hence, a shift of detectors to lower layers implies the presence of comparatively simple features specific to a category may influence participants reaction time or other behavioral metrics.

References

[1] K. Malakhova, *Category selectivity in neural networks*, https://github.com/taneta/category_selectivity_cnn, 2017.

[2] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, “Plug & play generative networks: Conditional iterative generation of images in latent space,” in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, IEEE, 2017, pp. 3510–3520.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, “Deep neural networks rival the representation of primate it cortex for core visual object recognition,” *PLoS computational biology*, vol. 10, no. 12, e1003963, 2014.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *ArXiv preprint arXiv:1409.1556*, 2014.

[6] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *ArXiv preprint arXiv:1312.6199*, 2013.