NeuroImage 152 (2017) 184–194

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Seeing it all: Convolutional network layers map the function of the human visual system

Michael Eickenberg[a,c,d,*], Alexandre Gramfort[b,c], Gaël Varoquaux[a,c], Bertrand Thirion[a,c]

[a] Inria Parietal Team, Inria Saclay, France
[b] CNRS LTCI, Télécom ParisTech, Université Paris-Saclay, France
[c] Neurospin, I2BM, DSV, CEA Saclay, France
[d] DATA Team, Informatics Department, Ecole normale supérieure, Paris, France

ABSTRACT

Convolutional networks used for computer vision represent candidate models for the computations performed in mammalian visual systems. We use them as a detailed model of human brain activity during the viewing of natural images by constructing predictive models based on their different layers and BOLD fMRI activations. Analyzing the predictive performance across layers yields characteristic fingerprints for each visual brain region: early visual areas are better described by lower level convolutional net layers and later visual areas by higher level net layers, exhibiting a progression across ventral and dorsal streams. Our predictive model generalizes beyond brain responses to natural images. We illustrate this on two experiments, namely retinotopy and face-place oppositions, by synthesizing brain activity and performing classical brain mapping upon it. The synthesis recovers the activations observed in the corresponding fMRI studies, showing that this deep encoding model captures representations of brain function that are universal across experimental paradigms.

1. Introduction

Human and primate visual systems are highly performant in recognizing objects and scenes, providing the basis of an excellent understanding of the ambient 3D world. The visual cortex is hierarchically organized, which means that many functional modules have feedforward and feedback connections compatible with a global ordering from lower levels to higher levels (Felleman and Van Essen, 1991). The concept of visual "pathways" or "streams" (Mishkin and Ungerleider, 1982; Goodale and Milner, 1992) is an established pattern which identifies principal directions of information flow for specific tasks, namely object representation in the "ventral stream" (from occipital cortex into temporal cortex) and localization and spatial computations in the "dorsal stream" (from occipital cortex into parietal cortex). They share much processing in the occipital early visual areas and less outside of them. The ventral visual stream encompasses visual areas V1, V2, V3, V4 and several inferotemporal (IT) regions. Pure feedforward pathways from V1 to IT (via other areas) exist, and probably account for rapid object recognition (Thorpe et al., 1996; Fabre-Thorpe et al., 2001).

Many parts of the human and primate visual cortices exhibit retinotopic organization in so-called visual field maps: The image presented to the retina is kept topographically intact in the next processing steps on the cortical surface (Wandell et al., 2007). This results in a one-to-one correspondence between a point on the retina and the "centers of processing" for that point in the visual field maps, such that neighboring points on the retina are processed nearby in the visual field maps as well.

The seminal work of Hubel and Wiesel (1959) showed that cat and other mammal V1 neurons selectively respond to edges with a certain location and orientation in the visual field.

This discovery inspired a long line of research investigating the nature of the computations performed in other visual regions and how they are implemented. As an example, certain monkey V2 neurons were found to react to combinations of orientations, such as corners Anzai et al. (2007). Recently, it has been put forward that V2 may be an efficient encoder of expected natural image statistics arising from interactions of first-order edges Freeman et al. (2013). V4 is reported to respond to more complex geometric shapes, color, and a large number of other stimulus characteristics. Recently it has been posited that V4 performs mid-level feature extraction towards the goal of bottom-up and top-down figure-ground segmentation (Roe et al., 2012). Further down the ventral pathway, neurons in the IT cortex have been shown to be selective to parts of objects, objects and faces

* Corresponding author at: DATA Team, Informatics Department, Ecole normale supérieure, Paris, France.
E-mail address: michael.eickenberg@nsup.org (M. Eickenberg).

http://dx.doi.org/10.1016/j.neuroimage.2016.10.001
Received 25 January 2016; Accepted 1 October 2016
Available online 21 October 2016

(Desimone et al., 1984; Logothetis et al., 1995). Taken together, these findings indicate an increasing trend in abstractness of the representations formed along the ventral stream.

FMRI has been used very successfully to identify and delineate the aforementioned visual field maps as well as brain regions that seem to specialize in certain tasks in the sense that their responses are particularly strong for specific types of stimuli. This type of result has typically been derived using statistical contrast maps opposing various visual stimuli. The contributions (Kanwisher et al., 1997; Downing et al., 2001; Epstein and Kanwisher, 1998), for instance, use this technique to localize specialized regions: areas for faces, body parts, places. Finer models, known as "encoding" models or forward modeling techniques (Naselaris et al., 2011), have been used to study the brain response to stimuli in greater detail (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011). This setting usually relies on richer models, going beyond binary contrasts, towards a more open description of the link between stimulus and activation. The validity of the corresponding stimulus representation is then established by testing how well it predicts brain activity, often with a linear model, by using cross-validation on held-out data.

For example, in Kay et al. (2008), almost 2000 naturalistic images were used as stimuli and the BOLD signal responses were then fit using a predictive model based on Gabor filterbank responses of the images shown. Primary visual cortex was very well modeled, but also extrastriate areas such as visual area V4 were well explained by the Gabor filter model.

In this contribution, we make use of the hierarchical organization of modern convolutional networks for object recognition to model human brain activity. We create encoding models (Naselaris et al., 2011) from the processing layers of the convolutional network OverFeat (Sermanet et al., 2013), which each represent feature maps at different levels of complexity. We train a linear predictive model of brain activity for each of the layers on the datasets of Kay et al. (2008) and Huth et al. (2012) and compare their ability to describe brain activity for every voxel by evaluating the predictive score on held-out data.

The scores of the different layers outline continuous progression profiles that are distinct in each visual area. We demonstrate that the model captures the cognitive architecture of the visual system by investigating its generalization capacity to vision-neuroscience paradigms beyond natural-image viewing. To do so we use stimuli unseen by our model, of which some come from totally different experiments and follow vastly different pixel statistics. Our predictive model, which can be seen as data-driven forward model to generate fMRI activations, is used to synthesize putative brain activation maps corresponding to these novel stimuli. This methodology enables our model to reproduce classical experiments in the extensive literature of paradigm-driven fMRI research. We consider two of these experiments: retinotopic mapping, i.e. the capturing of spatial information to sufficient accuracy for the generation of visual field maps, and a faces/places contrast to capture high-level information.

Previous work has used convolutional networks with fMRI data (Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014). However it focused on specific experiments. Showing that results generalize across datasets and paradigms brings an important novel step to the use of convolutional networks for the study of human vision. First, we show the validity of the approach on a new dataset with videos rather than still images. Second, we synthesize plausible brain activity to new images from completely different experiments that rely on hand-crafted, well controlled stimuli. These results demonstrate that convolutional networks capture universal representations of the stimuli that linearly map to and separate cognitive processes, such that this link generalizes to unseen experimental paradigms.

2. Biological relevance of multi-layer vision models

The Gabor filter pyramid employed in the original work of Kay et al. (2008) can be seen as an instance of a biologically inspired computer vision model. Indeed, all of modern computer vision, in its roots, has been inspired by biological vision. The basic filter extraction techniques at the beginning of the most successful computer vision pipelines are based on local image gradients or laplacians (Canny, 1986; Simoncelli and Freeman, 1995), which are operations that have been found in V1 as edge detection and in the LGN as center-surround features. The HMAX model was constructed to incorporate the idea of hierarchies of layers (Riesenhuber and Poggio, 1999). HMAX models are layered architectures that typically begin with edge detection using oriented filters, followed by a spatial and across-channel max-pooling. Subsequent layers implement other forms of localized (convolutional) processing, such as linear template matching. Using a supervised classifier at the end of this processing, it reached near state-of-the-art object recognition capacities in Serre et al. (2007).

The natural question to ask in the context of predictive modeling of BOLD fMRI in visual areas is "*What comes after the Gabor filter pyramid?*". The scattering transform model (Mallat, 2012; Bruna and Mallat, 2013) provided only one supplementary layer of which one cannot state much more than the existence of brain voxels which it models well (Eickenberg et al., 2013). The scattering transform is a cascade of complex wavelets and complex moduli, which has good mathematical stability properties and yields rich representations. The layers C1 and C2 of HMAX as used in Serre et al. (2007) were obtained using random templates taken from the preceding pooling layer activation. They were not geared optimally towards object recognition. This made the difference between layers difficult to evaluate (see e.g. Kriegeskorte et al., 2008). Although quite similar in architecture, deep artificial neural networks are of much greater interest here. Indeed, they optimize intermediate layers towards increasing overall performance in object detection, which is known to be performed also in IT cortex in humans and primates (see Cadieu et al. (2014) and Kriegeskorte et al. (2008)).

Artificial neural networks for computer vision attain state-of-the-art results with optimized feature hierarchies in a layered architecture composed of stacked layers with units that compute a linear transformation of the activations of previous layers followed by a simple pointwise nonlinearity. For instance, the first linear transformations are typically similar to Gabor filters and the corresponding nonlinearities perform edge detection. Recent breakthroughs in the field of artificial neural networks have led to a series of unprecedented improvements in a variety of tasks, all achieved with the same family of architectures. Notably in domains previously considered to be the strongholds of human superiority over machines, such as object and speech recognition, these algorithms have gained ground, and, under certain metrics, have surpassed human performance (LeCun et al., 2015).

Bridging to neuroscience, Cadieu et al. (2014) and Yamins et al. (2014), using electrophysiological data, have shown that IT neuron activity is predictive of object category in a similar way as the penultimate layer of a deep convolutional network which was not trained on the stimuli. Even more striking: a deep convolutional network can predict the activity of IT neurons much better than either lower-level computer vision models or object category predictors. Furthermore, deep convolutional networks trained on object categories and linked to neural activity with simple linear models predict this neural activity as well as the same network trained directly on neural data, suggesting that the encoding of object categories in the network is a good proxy for the representation of neural activity. These two works inspired us to investigate the link between computer-vision convolutional networks and brain activity with fMRI in order to obtain a global view of the system. Indeed, fMRI is much more noisy and indirect than electrophysiological data, but it brings a wide coverage of the visual system.

Inspection of the first layer of a convolutional net reveals that it is composed of filters strongly resembling Gabor filters, as well as color