Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Eight open questions in the computational modeling of higher sensory cortex

Daniel LK Yamins and James J DiCarlo

Propelled by advances in biologically inspired computer vision and artificial intelligence, the past five years have seen significant progress in using deep neural networks to model response patterns of neurons in visual cortex. In this paper, we briefly review this progress and then discuss eight key 'open questions' that we believe will drive research in computational models of sensory systems over the next five years, both in visual cortex and beyond.

Address
Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Yamins, Daniel LK (yamins@mit.edu)

Current Opinion in Neurobiology 2016, **37**:114–120

This review comes from a themed issue on **Neurobiology of cognitive behavior**

Edited by **Alla Karpova** and **Roozbeh Kiani**

For a complete overview see the Issue and the Editorial

Available online 26th February 2016

http://dx.doi.org/10.1016/j.conb.2016.02.001

0959-4388/© 2016 Elsevier Ltd. All rights reserved.

Any scientific development of long-term value opens up as many new questions as it answers. This is certainly the case with recent progress in building deep neural network models of visual cortex. In this piece, our goal is to briefly describe these recent advances, and to outline what we consider to be the most interesting open problems in cortical modeling, both in vision and beyond. We focus is on questions that will require both cutting-edge algorithmic developments as well as next-generation neuroscience and cognitive science experiments.

Brief review of recent progress

Starting with the seminal ideas of Hubel and Wiesel, work in visual systems neuroscience over the past 60 years has shown that the ventral visual stream generates invariant object recognition behavior via a hierarchically-organized series of cortical areas that encode object properties with increasing selectivity and tolerance [1•,2–5]. Early visual areas, such as V1 cortex, capture low-level features such as edges and center-surround patterns [6•,7]. In contrast, neural population responses in the highest ventral visual areas, inferior temporal (IT) cortex, can be used to decode object category, robust to significant variations present in natural images [8–10]. The featural content of mid-level visual areas such as V2, V3, and V4 is less well understood, but these areas appear to contain intermediate computations between simple edges and complex objects, along a pipeline of increasing receptive field sizes [1•,11–18].

Many of these observations can be captured mathematically via class of computational architectures known as Hierarchical Convolutional Neural Networks (HCNNs), a generalization of Hubel and Wiesel's simple and complex cells that has been developed over the past 30 years [19••,20•]. HCNN models are composed of several retinotopic layers combined in series. Each layer is very simple, but together they produce a deep, complex transformation of the input data—in theory, like the transformation produced in the ventral stream. However, mapping a single HCNN model to ventral stream neural data has proven extremely challenging [12], in part because subtle parameter changes (e.g. number of layers, local receptive field sizes, &c) can dramatically affect a model's match to neural data [21,22]. Recent work in visual cortex seeks to go beyond this powerful but broad-stroke understanding to identify concrete predictive models of ventral cortex, and then use these models to gain insight inaccessible without large-scale computational precision.

A key aspect of this approach has been *performance-based* optimization, in which the parameters of a large multilayer neural networks are chosen to optimize the networks' performance on a high-level, ecologically valid visual task [23•]. Leveraging computer vision and machine learning techniques, together with large amounts of real-world labelled images used as supervised training data [24••,25,26•], HCNNs have been produced that achieve near-human-level performance on challenging object categorization tasks [27].

Intriguingly, even though these networks are not directly optimized to fit neural data, their top hidden layers are nonetheless highly predictive of single-site neural responses as well population-level representations in IT cortex both in electrophysiological [23•,28], and fMRI data [29•,30]. Specifically, model units from the highest hidden layers of these performance-optimized HCNN can be linearly combined to produce synthetic 'neurons' that predict the image-by-image response patterns of sites in IT cortex. Moreover, the population of these synthetic neurons closely matches the representational dissimilarity

matrices (RDMs, [31]) of the macaque and human IT populations. These deep, performance-optimized neural networks have thus yielded the first quantitatively accurate, predictive model of the IT population response.

Moreover, high-throughput computational experiments evaluating thousands of HCNN models on both task performance and neural-predictivity metrics, have found a strong correlation between performance of high-level object recognition tasks and ability to explain IT cortical spiking data [23•]. The predictive power of these models is driven not just by categorization performance alone, as ideal observer models with perfect access to object identity do not themselves predict IT neural response patterns nearly as well as the hierarchical neural network units [23•].

Critically, these HCNN models are *mappable* not only to IT, but also to other levels of the ventral visual stream. Lower model layer filter weights resemble Gabor wavelets and are effective models of fMRI voxel responses in V1 voxel data [29•,30]. Along the same lines, intermediate HCNN layers are predictive of neural responses in V4 cortex [23•]. In other words: combining two general biological constraints—the behavioral constraint of recognition performance, and the architectural constraint imposed by the HCNN model class—leads to greatly improved models of multiple layers of the visual sensory cascade. An additional benefit of this approach is that each layer of the HCNN is a *basis set* for its corresponding cortical area, from which large numbers of IT-, V4- or V1-like units can be generated. A common assumption in visual neuroscience is that understanding the qualitative structure of tuning curves in lower cortical areas (e.g. gabor conjunctions in V2 or curvature in V4 [32]) is a necessary precursor to explaining higher visual cortex. These recent results show that higher-level constraints can yield quantitative models even when bottom-up primitives have not yet been identified.

The mapping between neural networks and cortical neural responses is still far from perfect. However, these recent results are encouraging, and they advance the understanding of the ventral stream in at least two new ways. First, the predictive accuracy of these models suggests that the principles of cortical processing may be best described at the level of architectural statistics (rather than precise wiring patterns), learning rules (rather than descriptors of tuning curves), and ethological task goals (rather than information transmission). Second, because models derived from this approach are both accurately predictive and generative, they act as hypothesis generators that can be richly interrogated to explore key open questions and enable the rational design of neuroscience experiments to answer those questions. Below we list eight exciting open questions that are now approachable from this new vantage point.

Why is IT cortex heterogenous at large spatial scales?

IT cortex is not a single monolithic computational mass in which output features are randomly intermixed across the cortical surface, but instead is likely to contain multiple retinotopic areas, with posterior IT, central IT, and anterior IT areas performing potentially different computations [1•]. It is also now known that specialized face, place, body, and color-preferring regions at the multiple-millimeter scale are found in each of these IT areas [33–36]. Are these the only regions? If so, why these and not others? How do the regions arise in the first place? Understanding this heterogeneity with computational models has two components: first, identifying whether and how the observed distributions of unit selectivities arise, independently of their spatial clustering; and second, explaining the observed spatial clustering.

Existing HCNN models could likely be used to generate detailed predictions about the unit distributions. A basic question is: to what extent are the existence of apparently specialized populations of units (e.g. face-selective units) strongly dependent on the semantic content of the training data of the neural networks? Will standard neural network model approaches yield observed unit populations if trained on datasets with a mix of semantic content close to that experienced by humans during development (e.g. a large fraction of faces)? How sensitive are unit selectivity distributions to this semantic content?

The second question, about spatial clustering, will require a more substantial extension of the HCNN framework, since those models make no specific predictions about how their units are to be mapped to the two-dimensional cortical sheet. It is possible that using a simple self-organizing map approach [37] to cluster in space units with similar feature tunings would explain a large fraction of the spatial structure in IT. However, there is some evidence that clustering may not be along purely geometric or featural lines—for example, body-preferring patches arise near face-preferring patches even though there is no obvious geometric similarity between these two categories [38]. If the known regions do not emerge in these types of models, it will be important to understand what additional principles are required to build them. If they do, it will also be of interest to search for new model-predicted regions that could subsequently be confirmed or falsified using primate fMRI and electrophysiolgy experiments.

Which visual properties are explicitly encoded in intermediate ventral stream areas?

Intermediate visual areas such as V2 and V4 have proven especially hard to understand because, unlike V1 and IT, they are removed both from low-level image

Download English Version:

https://daneshyari.com/en/article/6266177

Download Persian Version:

https://daneshyari.com/article/6266177

Daneshyari.com