Neural Networks 87 (2017) 8–21

Contents lists available at ScienceDirect

Neural Networks

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

An empirical model of activity in macaque inferior temporal cortex

Salman Khan *, Bryan Tripp

Department of Systems Design Engineering, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1
Center for Theoretical Neuroscience, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1

ARTICLE INFO

Article history:
Received 13 April 2016
Received in revised form 28 November 2016
Accepted 2 December 2016
Available online 13 December 2016

Keywords:
Inferior temporal cortex
Primate vision
Tuning curves
Empirical model

ABSTRACT

There are compelling computational models of many properties of the primate ventral visual stream, but a gap remains between the models and the physiology. To facilitate ongoing refinement of these models, we have compiled diverse information from the electrophysiology literature into a statistical model of inferotemporal (IT) cortex responses. This is a purely descriptive model, so it has little explanatory power. However it is able to directly incorporate a rich and extensible set of tuning properties. So far, we have approximated tuning curves and statistics of tuning diversity for occlusion, clutter, size, orientation, position, and object selectivity in early versus late response phases. We integrated the model with the V-REP simulator, which provides stimulus properties in a simulated physical environment. In contrast with the empirical model presented here, mechanistic models are ultimately more useful for understanding neural systems. However, a detailed empirical model may be useful as a source of labeled data for optimizing and validating mechanistic models, or as a source of input to models of other brain areas.

1. Introduction

The primate inferotemporal (IT) cortex is a high-level visual area in the ventral visual stream. Its neurons respond strongly to complex visual features, and these responses exhibit tolerances to many stimulus transformations. IT has strong connections with areas around the hippocampus, and also with the prefrontal cortex (Webster, Bachevalier, & Ungerleider, 1994). IT provides information for object recognition (e.g. Hung, Kreiman, Poggio, & DiCarlo, 2005; Tanaka, 1996). Characterization of the visual representation in IT is important for understanding the mechanisms through which this representation emerges from lower-level visual features, and also for understanding the visual signals available to other areas.

As a step in understanding the visual representation in IT in more detail, we sought to combine a wide range of information from the literature into a coherent statistical model of IT activity. In contrast with network models of IT (Rolls, 2012; Serre, Kreiman et al., 2007), which offer mechanistic explanations for IT tuning properties, we developed a purely descriptive model. Its advantages are that it incorporates a wide range of response properties, and that it can be extended in a straightforward way to incorporate more response properties as needed. On the other hand, since it does not incorporate realistic mechanisms, it is less likely than a realistic mechanistic model to extrapolate accurately beyond tuning properties that are not explicitly incorporated.

Given this limitation, the main intended use of the model is to support development of future mechanistic models. In particular, we expect that it can serve as a source of labeled data with which to refine mechanistic ventral-stream models. Previous mechanistic models of the ventral stream have already been quite successful, but we expect that further improvements may be facilitated by taking a step back and more thoroughly modeling IT response statistics, as we do here. However, we also emphasize that although we have tried to incorporate diverse data into the model, we have so far only been able to address a small fraction of the IT literature.

1.1. Previous IT models

The best-known previous models of the inferotemporal cortex (IT) include the Neocognitron (Fukushima, 1980), HMAX (Serre, Kreiman et al., 2007), and VisNet (Rolls, 2012). Convolutional networks trained for object categorization have also been considered as IT models (Hong, Yamins, Majaj, & DiCarlo, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Here we briefly describe these models and their limitations.

The HMAX model, a successor of the Neocognitron model, is composed of a hierarchical series of alternating 'simple' and

* Corresponding author at: Department of Systems Design Engineering, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1.
E-mail addresses: s362khan@uwaterloo.ca (S. Khan), bptripp@uwaterloo.ca (B. Tripp).

http://dx.doi.org/10.1016/j.neunet.2016.12.001
0893-6080/

'complex' cell layers designed to model rapid feed-forward visual processing. Simple cells extract features, while complex cells pool extracted features to develop tolerances to both spatial and scale transformations. With increasing depth, both the complexity of the features and tolerances to these transformations increase. HMAX is consistent with many properties of IT responses, such as orientation and clutter tuning (Serre et al., 2005). It performs simple object-recognition tasks at human performance levels (Serre, Oliva, & Poggio, 2007). Limitations of HMAX have been discussed extensively by Robinson and Rolls (2015) (for example, it responds similarly to scrambled and unscrambled images of faces). Further limitations include that HMAX activity does not reflect category distinctions in the same way as IT activity (Khaligh-Razavi & Kriegeskorte, 2014), and that its performance in challenging object classification tasks is poor (Cadieu et al., 2014).

VisNet is also a well-established IT model that has been extensively developed and compared with IT (Robinson & Rolls, 2015; Rolls, 2012). It uses a learning rule with a memory trace that has the effect of correlating responses to stimuli that are presented close together in time. This can introduce realistic invariances to position, scale, and rotation, depending on the pattern of stimulus presentations. It is argued that (for example) position-invariant responses to a certain object could emerge naturally as an object moves across the visual field, and size-invariant responses could emerge as an object recedes from the viewer. Position, size, and rotation invariance have all been established in VisNet models (Rolls, 2012). However, a limitation is that different stimulation protocols have been used in each case, and these tests have not involved naturalistic scenes and eye-movement patterns. VisNet does not produce responses that are realistically correlated across categories (Khaligh-Razavi & Kriegeskorte, 2014). Its object categorization performance is also comparable to that of HMAX (Robinson & Rolls, 2015), far below human performance.

Deep convolutional neural networks (LeCun, Bottou, Bengio, & Haffner, 1998) are artificial neural networks with sparse, local connections. They were loosely inspired by the visual cortex, and have since demonstrated human-level performance in core object recognition (He, Zhang, Ren, & Sun, 2015). They are physiologically unrealistic in a number of ways. For example, they are typically trained with backpropagation, and their weight updates are non-local, in that weights are shared across feature maps. (This weight sharing is sometimes relaxed or eliminated (Bishop, 2006), but it is very useful for making efficient use of available training data. This is a reasonable accommodation, since the networks are trained with fairly small datasets compared to the enormous volume of visual data available in life.) They have other unrealistic properties, e.g. highly simplified neurons, that are shared by VisNet and HMAX.

Despite these simplifications, convolutional neural networks (CNNs) are currently the only models that approach human performance in rapid object categorization in natural scenes, a function closely related to IT. Interestingly, convolutional networks trained for object classification also represent object position and orientation in their later layers, much like IT neurons (Hong et al., 2016). Also like IT (and unlike HMAX or VisNet) their internal activity is correlated within object categories (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Furthermore, internal activity in their later layers predicts much of the variance in activity of real IT neurons in response to the same stimuli (Cadieu et al., 2014; Yamins et al., 2014), and internal activity in their intermediate layers similarly predicts neural activity in V4 (Yamins et al., 2014). Their internal activity is also more invariant to position, rotation, and scale in later layers than earlier layers, analogous to differences between V1 and IT (Zeiler & Fergus, 2014).

In light of these similarities, we examined several additional properties of CNN unit responses and compared them with those of our empirical model. We found that CNN responses are strikingly similar to IT responses in many respects. Some CNN properties, including the distribution of population sparseness across stimuli, and the distribution of size bandwidths, were essentially indistinguishable from corresponding IT properties, and other properties differed to various degrees. However, the substantial similarities suggest that further increasing the similarity between CNNs and IT may yield a model with responses that are hard to distinguish from early IT responses, and we hope that our empirical model can contribute to this convergence.

In summary, a variety of previous models have approximated various properties of IT, and provided various insights, but all of them have activity patterns that are currently easily distinguishable from those of IT. However, the gap between CNNs and the earliest (largely feedforward) responses in IT is fairly small. A large part of our motivation for developing the present model is to provide labels for training CNNs in a way that further reduces this gap (discussed further in the Discussion).

2. Methods

The input to the model is a list of visible objects and corresponding parameters (e.g. retinal position, degree of occlusion, etc.) The output is a list of spike rates for a population of IT neurons. We sometimes additionally modeled spike-rate changes over time, and produced inhomogeneous Poisson spikes as output. The inputs can be specified manually, but we also developed an interface with the robotics simulator V-REP (Rohmer, Singh, & Freese, 2013). This interface calculates the relevant stimulus parameters from the simulation environment.

The complete simulation software (written mainly in Python) is freely available at https://github.com/salkhan23/ITCortex.

Many of the statistical distributions in the model are fits to data for which distributions have not been proposed previously. We used existing models where possible. We modified one published model (the distribution of object selectivities), because this was necessary in order to integrate this distribution with the rest of our model.

2.1. Model structure and data sources

A neuron's spike rate in response to a single stimulus object was approximated as the product of an object-dependent "unscaled" spike rate r_{obj}, and various scale factors $0 \leq s \leq 1$. Specifically, a given neuron's spike rate response to a single *isolated* stimulus object was,

$$r_{iso} = r_{obj} s_{pos} s_{size} s_{rot} s_{occ}, \quad (1)$$

where s_{pos}, s_{size}, s_{rot}, and s_{occ} are taken from the neuron's tuning curves for retinal position, size, rotation, and occlusion, respectively. We modeled both the forms and parameter distributions of the tuning curves on the IT electrophysiology literature, as described in detail in the following sections. We also modeled effects of clutter and dynamics, as described in later sections.

Where possible, we used data from studies with large numbers of cells, and with information about the statistics of response distributions. In the absence of evidence to the contrary, we assumed that tuning for multiple stimulus properties was separable, i.e. that we could approximate effects of multiple parameters as a product of scaling factors. Examples of stimulus and position preference in Figure 5 of DiCarlo and Maunsell (2003) are generally consistent with this assumption. Similarly, Op De Beeck and Vogels (2000) reported that position sensitivity was invariant to changes in stimulus shape and size. The only counter-example in our model is occlusion of diagnostic and non-diagnostic object parts (see Section 2.3.4), which we modeled as a non-separable two-dimensional