Improving understanding of intermediate and higher cortical areas by training models in virtual worlds

Chengxu Zhuang[1*], Damian Mrowca[2], Daniel Yamins[1,2,3]

[1]Psychology Department, Stanford University
[2]Computer Science Department, Stanford University
[3]Stanford Neuroscience Institute
Jordan Hall Rm 427, Stanford, CA 94305, USA

[*]To whom correspondence should be addressed; E-mail: chengxuz@stanford.edu.

In this paper, we explore the use of virtual three-dimensional worlds to train deep neural networks to solve object-related tasks, both at high and intermediate levels of representation. Specifically, we consider two sensory domains: (1) a typical visual domain, in which the neural network is exposed to visual images created by rendering in a complex three-dimensional scene, and (2) a whisker-like domain in which a vibrissal array is passed across 3-d objects in these same types of scenes, and which the neural network is driven by the array of forces at the base of the vibrissae. In both cases of input format, we train the networks to solve object-related tasks, including (1) object category recognition, a task putatively engaging high-level abstractions and (2) agent-centered object normal map estimation, a task that is putatively at an intermediate level of abstraction. After training these models, we then investigate how well they predict responses in neurons in animal brain areas believed to underly

the above-mentioned recognition behaviors. Specifically, for the visual input case, we seek to determine whether networks trained on both intermediate and high-level tasks lead are improved models of the macaque and human ventral visual stream, compared to existing Deep Neural Networks trained only on high-level tasks. For the whisker input case, we seek to determine whether any of the networks so trained are effective models of higher barrel cortex in rodents, including areas S1 and S2. We also hope that this work will inspire future researchers to utilize the considerable literature about intermediate and high-level behavioral functions to improve predict power of models both in solving sensory tasks and in modeling brains.

Introduction

Brains do remarkable work in actively analyzing environmental information and making decisions based on this information. Among all brain systems, sensory systems are especially concentrated on analyzing some particular environmental information, e.g., light for visual systems, sound for auditory systems, and inputs from whiskers for the rodent somatosensory systems [25]. The goal of these sensory systems is to extract behaviorally useful information from the complex raw input data, a process which could be described as untangling the behavior-related dimensions (such as category) from other orthogonal dimensions (such as translation and rotation of the objects)[36]. In this work, we are especially interested in visual systems and the rodent somatosensory systems.

While these two systems radically differ in their input modalities, number of overall neurons, and specific neuronal microcircuits, they share two fundamental characteristics. First, both are **deep sensory cascades**, consisting of sequences of brain areas (cortical and subcortical), each of which alone is comparatively simple, but which together produce a complex transformation of the input data. Second both are inherently spatially extended (in contrast to, for instance, the olfactory sense, where any fundamental spatial structure on the input is much less clear) [9]. In this work, we leverage these similarities to produce models for both that share some core common architectural principles.

Starting from the seminal work of Hubel and Wiesel, researchers have made significant progress in understanding the primate visual systems. And for the ventral stream [10] which we are especially interested in, different areas have also been characterized in different functions. For example, V1 (primary visual cortex) has been described as detecting lower level features of images, such as edges and colors, while V2 is found to be able to detect more complicated features as contours. The function of V4 and IT is not well understood, but they are responsible for detecting much more complicated features than in V2, such as shapes for V4 and object categories for IT.

In somatosensory systems, researchers have also found evidence showing hierarchical processing for somatosensory input in rodents, human, and primates[24, 16, 17]. For example, the second somatosensory area (S2) is found to rely on inputs from S1 (first somatosensory area) [24, 23]. And barrel cortex, S1, and S2 are also found to get different aspects of input from rodent whiskers [7]. And connections between barrel cortex and S1, S1 and S2 are also believed to convey information for hierarchical processing [23]. While barrel cortex is already explored a lot, the functions of both S1 and S2 is poorly understood.

A fundamental question that arises from these experimental observations is: what are the underlying algorithms behind these areas? What computations are they doing? These questions naturally take the form of **computational neuroscience**: e.g. producing a computational model that correctly describes neural responses as a function of stimulus input.

Hierarchical models are also used widely in artificial intelligence to help design better systems for various tasks. Recent work using deep neural networks (DNNs) has achieved significant improvements on object recognition, speech recognition, and numerous other artificial intelligence tasks[19, 13, 21]. These deep neural networks are all composed of multiple simple neural network layers in series, where the computation in single layer is usually simple but non-linear and stacks of those simple non-linear computations make up a highly complicated non-linear computation. DNNs are also believed to be biologically plausible, and therefore could be good candidates for models of sensory cortical brain systems.

In fact, researchers have also found that the DNNs optimized for performances on object recognition tasks serve as a good model for the primate ventral visual stream [34, 35, 5]. Using a general class of computational architectures known as HCNNs [20], which also include the recent convolutional neural networks, they found that optimizing for object recognition task could improve not only the performances on explaining the responses of IT layers using that of final layers but also that of explaining V4 areas using intermediate layers of the same architecture simultaneously [5]. And the final models they got successfully surpassed all existed models on explaining IT and V4 areas. Similar results were also found in the deep neural networks that were trained on large-scale object recognition tasks [35].

However, even the best existing DNN models still capture V4 and IT imperfectly. V4 is believed to encode intermediate level of object features and show strong attentional modulation[28], while IT areas are believed to encode high level of object features as object category.

Our hypothesis in this work is that by taking into account this intermediate task of normal encoding at an intermediate model level during training, we will produce a better model both of V4 and of IT. Specifically, then, the task that we are interested in for V4 neurons is predicting normals of object surfaces using 2D images as input. Once we have a good model for V4, we could further add some extra layers on top of it to predict the object category from normals. Those extra layers then could be treated as models for IT areas.

For the rodent somatosensory cortex, we are using a similar task but with input now collected through simulated-whiskers to simulate the input to mouse barrel cortex. Rodents could use their whiskers to detect object shape, position, and texture of object surface[4, 7, 3, 22]. And our hypothesis in explaining this is that area S1 (primary somatosensory cortex) in the rodent is a normal estimator, while area S2 (secondary somatosensory cortex) is a higher-level object category or shape detector. By explicitly modeling the functions of intermediate layers, we hope that we could have a better model for the whole systems. In fact, unlike the case of visual cortex, where there are many existing models, in rodent barrel cortex, there is significantly less modeling work, and we hope our contribution will be the first substantial such model for S1 and S2.

A core problem in carrying out the plan described above for optimizing deep nets on intermediate-level tasks is the lack of sufficient annotated training data from natural real-world stimuli. For the visual case, it is extremely challenging to collect high-quality normal estimations of sufficiently many natural scenes that training deep nets on such data is easily effective. Similarly, it is very challenging to obtain directly what whiskers in a mouse experience, or to build a real-world sensory device mimicking rodent whisker arrays. We were therefore motivated to create virtual worlds in which these data would be readily available, and train our networks in those virtual worlds. [You should cite existing work (Eigen and Fergus, for example) on training normal extraction networks.]

There has already been some promising work investigating training deep hierarchical models in virtual worlds[26, 18]. In the work by Johnson-Roberson and et al.[18], virtual worlds have been used to train deep hierarchical models to learn driving. And in our work, we will illustrate that it is possible that we could successfully train a model that performs well on object recognition task both in virtual environment and real world without using any training examples and notations from real world.

As for building a virtual model of the rodent whisker input device, we will make use of recent investigations into the rodent whisker and barrel cortex[4, 7, 3, 22]. Mechanical properties of the rodent whiskers have also been explored in detail [7, 27, 30]. Recently, researchers have begun to utilize the known mechanical properties of vibrissae to simulate the responses of whiskers under particular circumstances [15]. Combining these insights with deep hierarchical models in a virtual world, we will build a mode for both barrel cortex and mouse somatosensory systems.

In the following sections, we will describe the methods we use for building the two virtual worlds and designing/training the neural networks, describe the results obtained both for macaque visual and mouse barrel cortex, and then briefly discuss the implications of these results.

Methods

In this section, we will explain how we build the virtual worlds and then how we train models using them.

Virtual worlds

We build two different virtual worlds for the two systems respectively. For V4 and IT areas modelling, the virtual world we want should be able to generate realistic images showing different objects in various scenes as well as providing the needed notations, which, in our case, is normals on object surfaces and categories of those objects. In order to make our models trained in this virtual world generalize well to real worlds, we use one of the most popular and realistic game engines, Unity 5 [2], to build our virtual world, which we refer to as 3Dworld. Besides Unity, we also need plenty of realistic 3D objects to be placed in the virtual world. We collected the models from various sources, including Shapenet [6], Dosch 3D models [1]. To generate 2D images and related notations for training the models, we will first build a scene in this virtual world with reasonable light condition and semantic surroundings (for example, walls and windows for indoor scenes or grass and trees for outdoor scenes). Following that, we will place different 3D objects in the scene. Currently, the 3D objects are chosen randomly from the whole library of 3D objects we have. Then, we will randomly take images including reasonable number of 3D objects that occupy a reasonable part of the whole image. Finally, the images and normals of objects shown in the image as well as their categories will be used as training examples for deep hierarchical models.

For mouse somatosensory systems, the virtual world should first have a physically accurate models for the the rodent whiskers. Researchers in computer graphics have done great job modelling human hair or furry objects[12, 31], while their concentration is usually to make tens of thousands of hair strands have reasonable behaviors interacting with each other or with bodies and other objects. And in our simulations, we are more interested in ensuring the correctness of the mechanical details of every individual whisker. For this virtual world, we use Bullet [33] as physics engine and the 3D objects used are from the same library for the 3Dworld. Every unit is in shape of a cuboid and connected to two neighbors using hinge connections. Specifically, we use concatenated 25 small units in a chain to model the individual whisker. To simplify the discussion, we assume the whiskers are laid along X-axis in a 3D space with XYZ axis. All the hinge connections could only rotate in XOY plane. And the base unit is connected to a base ball representing the actual connecting point of whiskers to mouse. The connection between the base unit and base ball is also hinge connection which is only allowed to rotate in ZOY plane. With only those hinges between every two concatenated units as constraints, different parts of whisker strand would have independent actions as the constraints for two units in large distance would take a long time to actually work. Thus we also add springs between pairs of units that are not neighbors. For example, we add springs connecting every two units that have 2 other units between them. After adding all those constraints, we could then build an array of individual whisker strands to simulate the whisker array on mouse. And following that, we use the simulated whiskers to sweep over surfaces of objects and collect the responses of the whiskers by measuring how much the several hinges and springs near base unit leave their equilibrium states at every time point. Annotated with normals of the surfaces swept and categories of those objects, those data could be used to train the deep hierarchical models.

Model training

We could have various ways to train our models solving these tasks: (1) end-to-end training for the whole task, e.g. from pixel values to object category for visual system; (2) from input to intermediate target, and then from intermediate to final target, e.g., for the rodent somatosensory systems it is from simulated-whisker inputs to normals, and then from normals to object category; (3) doing both of these two training simultaneously.

We are also interested in comparing these training methods. By comparing their performances on tasks or performances on predicting neuron responses, we could answer whether the second and third training methods would be better than the first. More specifically for second training method, we also need to investigate the number of layers we would take to get the intermediate target and similarly, how many extra layers on the intermediate output we need to get the final target. As the constraint of getting the intermediate target could be a regularization to the whole network. We might expect that less training examples overall would be needed in second and third training methods compared to the first. And especially for the second stage of training (from intermediate target to final target), as the task should be easier than from raw input to final target, we would also expect that less training examples would be needed.

Those training methods can be applied to both two systems, as in both two systems, we are interested in predicting intermediate target and final target as well. But the models we will use need to be different, as the input data for models in two systems is very different.

For visual systems, our models should be able to estimate the normals or object category map from static pictures for the scenes. The possible network structures are that used in Eigen and Fergus[8], where the same multi-scale convolutional deep neural network was used for both predicting normals and categories.

As for networks for the rodent somatosensory systems, the input data changes from static images to time sequences of responses from arrays of simulated-whiskers. Therefore, our networks should be able to process the time series information and then estimate normals and category labels from that. We propose three ways to do this:

(1) concatenate all time sequences at the channel dimension. For example, if the responses from simulated-whisker array is in dimension of $W \times H \times M$, where W is width of array, H is height of array, and M is number of mechanical measures we have for every individual whisker, and we have T frames of responses, then we could concatenate those stimuli to be in dimension of $W \times H \times (MT)$. After that, we could train the usual convolutional neural networks on it predicting normals or categories.

(2) The problem about method (1) above is that with the response concatenating, we will either need a more larger number of parameters or be lack of complexity needed to capture the structure inside the stimuli. One advantage that we could have taken is that the different frames of responses have similar structure. And by concatenating them, we are ignoring this similarity and expecting that our model could discover that through learning, which therefore requires more complexity in the models. So to make use of this advantage, we could first have a smaller network for every frame of responses and instead of concatenating these frames of responses, we concatenate the output of those small networks for each frames and then train another network from them to our targets, whether intermediate or final.

(3) As the problem we are facing is very familiar as that for auditory cortex after converting sound signals to fourier space, we could apply results for acoustic speech recognition using

deep neural networks here as well [11, 29, 32]. The networks they used either used a special structure called LSTM [14] in traditional convolutional neural networks to equip them with dynamics across time dimention or used simple recurrent connections inside the same layer of network conveying information from one frame to another. We will investigate these possible structures in our problem to choose the one that provide best performance.

Results

We will explain this later, not in this proposal.

Discussion

We will explain this later, not in this proposal.

References

[1] Dosch homepage, https://www.doschdesign.com/.

[2] Unity homepage, https://unity3d.com/.

[3] Ehsan Arabzadeh, Erik Zorzin, and Mathew E. Diamond. Neuronal encoding of texture in the whisker sensory pathway. *PLoS Biology*, 3(1), 2005.

[4] Yves Boubenec, Daniel E Shulz, and Georges Debrégeas. Whisker encoding of mechanical events during active tactile exploration. *Front Behav Neurosci*, 6(November):74, 2012.

[5] Charles F. Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, 10(12):1–35, 2014.

[6] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. *ArXiv*, 2015.

[7] Mathew E Diamond, Moritz von Heimendahl, Per Magne Knutsen, David Kleinfeld, and Ehud Ahissar. 'Where' and 'what' in the whisker sensorimotor system. *Nat Rev Neurosci*, 9(8):601–612, 2008.

[8] David Eigen and Rob Fergus. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture. 2014.

[9] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47, 1991.

[10] Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

[11] A Graves, A.-R. Mohamed, and G Hinton. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (6):6645–6649, 2013.

[12] Sunil Hadap, Marie-Paule Cani, Ming Lin, Tae-Yong Kim, Florence Bertails, Steve Marschner, Kelly Ward, and Zoran Kači-Alesi. Realistic Hair Simulation Animation and Rendering. 2008.

[13] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[14] Sepp Hochreiter and Jurgen Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1–32, 1997.

[15] Lucie A. Huet and Mitra J Z Hartmann. Simulations of a Vibrissa Slipping along a S-traight Edge and an Analysis of Frictional Effects during Whisking. *IEEE Transactions on Haptics*, 9(2):158–169, 2016.

[16] Koji Inui, Xiaohong Wang, Yohei Tamura, Yoshiki Kaneoke, and Ryusuke Kakigi. Serial processing in the human somatosensory system. *Cerebral Cortex*, 14(8):851–857, 2004.

[17] Yoshiaki Iwamura. Hierarchical somatosensory processing. *Current Opinion in Neurobiology*, 8(4):522–528, 1998.

[18] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, and Ram Vasudevan. Driving in the Matrix: Can Virtual Worlds Replace Human-Generated Annotations for Real World Tasks? 2016.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[20] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[22] Daniel H. O'Connor, Simon P. Peron, Daniel Huber, and Karel Svoboda. Neural activity in barrel cortex underlying vibrissa-based object localization in mice. *Neuron*, 67(6):1048–1061, 2010.

[23] Carl C.H. Petersen. The Functional Organization of the Barrel Cortex. *Neuron*, 56(2):339–355, 2007.

[24] T P Pons, P E Garraghty, David P Friedman, and Mortimer Mishkin. Physiological evidence for serial processing in somatosensory cortex. *Science (New York, N.Y.)*, 237(4813):417–420, 1987.

[25] Dale Purves, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel LaMantia, James O McNamara, and S Mark Williams. Neuroscience. *Sunderland, MA: Sinauer Associates*, 3, 2001.

[26] Weichao Qiu and Alan Yuille. UnrealCV: Connecting Computer Vision to Unreal Engine. pages 1–8, 2016.

[27] Brian W Quist, Vlad Seghete, Lucie A Huet, Todd D Murphey, and Mitra J Z Hartmann. Modeling Forces and Moments at the Base of a Rat Vibrissa during Noncontact Whisking and Whisking against an Object. *J Neurosci*, 34(30):9828–9844, 2014.

[28] Anna W. Roe, Leonardo Chelazzi, Charles E. Connor, Bevil R. Conway, Ichiro Fujita, Jack L. Gallant, Haidong Lu, and Wim Vanduffel. Toward a Unified Theory of Visual Area V4. *Neuron*, 74(1):12–29, 2012.

[29] Hasim Sak, Andrew Senior, and Françoise Beaufays. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. *Interspeech 2014*, (September):338–342, 2014.

[30] R. Blythe Towal, Brian W. Quist, Venkatesh Gopal, Joseph H. Solomon, and Mitra J Z Hartmann. The morphology of the rat vibrissal array: A model for quantifying spatiotemporal patterns of whisker-object contact. *PLoS Computational Biology*, 7(4), 2011.

[31] Kelly Ward, Florence Bertails, Tae Yong Kim, Stephen R. Marschner, Marie Paule Cani, and Ming C. Lin. A survey on hair modeling: Styling, simulation, and rendering. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):213–233, 2007.

[32] Chao Weng, Dong Yu, Shinji Watanabe, and Biing-Hwang Fred Juang. Recurrent deep neural networks for robust speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, (2):5532–5536, 2014.

[33] Wikipedia. Bullet (software) — wikipedia, the free encyclopedia, 2016. [Online; accessed 19-October-2016].

[34] D L Yamins, H Hong, and C Cadieu. Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. *Advances in neural information processing systems*, pages 1–9, 2013.

[35] Daniel L K Yamins, Ha Hong, Charles F Cadieu, Ethan a Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24, jun 2014.

[36] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.