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]Neuroscience Institute, Stanford University
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 these models predict responses in neurons in animal brain areas believed

to underly these functions. 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 have done remarkable work by actively analyzing environment information and making decisions upon that. Among all systems in brains, sensory systems are usually concentrated on analyzing some particular environment information, for example, light for visual systems, sound for auditory systems, and inputs from whiskers for mouse somatosensory systems. The goal of those sensory systems is to extract the useful semantic information from the complex raw input data, which could be described as untangling the behavior-related dimensions (such as category) from other irrelevant dimensions (such as translation and rotation of the objects)[33]. And in this work, we are especially interested in visual systems and mouse somatosensory systems.

A lot of work has been done to explore both two systems. While those systems differ from their input data, number of overall neurons, and specific structures as well as organizations, they are believed to have the similarity of consisting of several consecutive regions that are distinguishable on both structures and functions [9]. For visual systems, starting from work

of Hubel and Wiesel[13], there have been a large number of hierarchical models developed to explain the response patterns of them[25, 27, 10, 4, 21]. Similarly in somatosensory systems, researchers also find evidence showing hierarchical processing for somatosensory input in both human and primates[22, 15, 16].

Hierarchical models are also used widely in artificial intelligence to help design better systems for various tasks. The recent work using deep neural networks (DNNs) has achieved significant improvements on object recognition, speech recognition, and numerous of other artificial intelligence tasks[18, 12, 19]. Those 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 finally make up of some highly complicated non-linear computations. Additionally, those models are also believed to be biologically plausible and therefore could be good candidates for models of related brain systems.

Furthermore, researchers have also found that the hierarchical models optimized for performances on object recognition tasks could also serve as a good model for IT areas in primates, which are believed to be the responsible areas for object recognition in brains [31, 32, 6]. Inspired by this, we are building performance optimized hierarchical models for specific tasks that we think V4 (an intermediate layer in visual systems) in human and primates and mouse somatosensory systems are performing. And after having the models, we could use them to explain the responses of those brain areas.

Both two areas of interest are poor understood. V4 is believed to encode intermediate level of object features and show strong attentional modulation[26], while IT areas are believed to encode high level of object features as object category. And mouse could use their whiskers to detect object shape, position, and texture of object surface[5, 8, 3, 20]. Thus 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 mouse somatosensory cortex, we are using the similar task but with input now collected through simulated-whiskers to simulate the input to mouse barrel cortex. Via doing this, we are trying to model S1 (primary somatosensory cortex) as normal predictors and then S2 (secondary somatosensory cortex) as object category detector. With explicitly modeling the functions of intermediate layers, we hope that we could have a better model for the whole systems.

However, for optimizing those deep hierarchical models, one would need a large nubmer of examples with corresponding labels. And in our work, it is either too difficult to collect the corresponding labels (for example, the normals of object surfaces given the 2D images) or the desired example itself (for example, the input from simulated whiskers). Thus we need to create virtual worlds for our tasks and train our models there.

There have already been some works investigating training deep hierarchical models in virtual worlds[23, 17]. In the work by Johnson-Roberson and et al.[17], 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 mouse somatosensory systems, a lot of investigations have been made to mouse whiskers and barrel cortex[5, 8, 3, 20]. Mechanical properties of mouse whiskers have also been explored a lot [8, 24, 28]. Recently, researchers started to utilize the mechanical properties known to simulate the responses of whiskers under particular circumstances [14]. Combining them 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 our methods used for both building the virtual worlds and training the models in Methods section. [TODO] Then in Results section, we will illustrate our results from the models and their performances on explaining those systems. [TODO] We will briefly discuss those results in Discuss section.

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 [7], 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 mouse whiskers. Researchers in computer graphics have done great job modelling human hair or furry objects[11, 29], 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 [30] 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 will explain this later, not in this proposal.

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] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

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

[6] 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.

[7] 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.

[8] 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.

[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] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[11] 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.

[12] 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.

[13] D H Hubel and T N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148:574–591, 1959.

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

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] Nicolas Pinto, David Doukhan, James J DiCarlo, and David D Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol*, 5(11):e1000579, 2009.

[22] 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.

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

[24] 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.

[25] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

[26] 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.

[27] Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429, 2007.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

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