frontiers in Computational Neuroscience

GENERAL COMMENTARY
published: 19 January 2018
doi: 10.3389/fncom.2018.00004

Commentary: Using goal-driven deep learning models to understand sensory cortex

*Qiulei Dong [1,2,3], Hong Wang [2] and Zhanyi Hu [1,2,3]**

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2] University of Chinese Academy of Sciences, Beijing, China, [3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

Keywords: goal-driven deep learning models, hierarchical convolutional neural network, IT neuron, categorization, convergent features

A commentary on

Using goal-driven deep learning models to understand sensory cortex
by Yamins, D., and DiCarlo, J. (2016). Nat. Neurosci. 19, 356–365. doi: 10.1038/nn.4244

OPEN ACCESS

Edited by:
Guenther Palm,
University of Ulm, Germany

Reviewed by:
Thomas Wennekers,
Plymouth University, United Kingdom

****Correspondence:***
Zhanyi Hu
huzy@nlpr.ia.ac.cn

Received: *13 September 2017*
Accepted: *05 January 2018*
Published: *19 January 2018*

Citation:
Dong Q, Wang H and Hu Z (2018) Commentary: Using goal-driven deep learning models to understand sensory cortex. Front. Comput. Neurosci. 12:4. doi: 10.3389/fncom.2018.00004

Recently, a goal-driven modeling approach of sensory cortex is proposed in Yamins and DiCarlo (2016). The basic idea of this approach is to first optimize a hierarchical convolutional neural network (HCNN) for performing an ethologically relevant task, then once the network parameters have been fixed, to compare the outputs of different layers of the network to neural data. The success of this approach is exemplified by the results in Yamins et al. (2014), where a 4-layer HCNN, called HMO, was used to predict IT neuron spikes on image object stimuli. Notably by only optimizing the 8-way image categorization performances, not only can the top output layer of the HMO quantitatively predict IT neuron responses, but its penultimate layer can also automatically predict V4 neuron responses. In Hong et al. (2016), under the same approach, a 6-layer HCNN was trained on ImageNet (Russakovsky et al., 2015) (a benchmark dataset for image object categorization in the computer vision field, containing 1.3 million category-labeled training images of 1,000 different categories) to successfully predict category-orthogonal object properties along the ventral stream. Another demonstrative example is the work in Khaligh-Razavi and Kriegeskorte (2014), showing that when the 10-category representational dissimilarity matrices were used together with the outputs of all the 8 layers of the AlexNet in Krizhevsky et al. (2012), called the IT-geometry supervised layer, its outputs could sufficiently explain IT data.

Here in this commentary, we would say that this goal-driven approach, although with some notable successes and great potential for understanding sensory cortex, could be not as general as the authors (Yamins and DiCarlo, 2016) advocate, and its general use should be taken with special care. This is because as shown in Li et al. (2016), the 4 different HCNNs, with the same AlexNet architecture trained with the same dataset (ImageNet) but only from different random initializations, learned both convergent and divergent features although the 4 HCNNs have achieved the similar categorization performances: their top-1 accuracies are 58.65, 58.73, 58.79, and 58.84% respectively, which are also similar to the top-1 performance of 59.3% reported in the original study (Krizhevsky et al., 2012). In other words, some convergent features, which are individually similar or related via a linear transformation, are reliably learnt by the 4 HCNNs, yet other divergent features are not consistently learnt. In particular, the features at downstream layers are more divergent than convergent among the 4 HCNNs. The divergence is particularly marked by two aspects: (1) The responses of neurons at higher layers in one network were impossible to be linearly mapped to the responses of the neurons at the same layer in other networks (Table 1 in Li et al., 2016). Or the outputs of the neurons at the same layer in a pair of networks cannot be adequately related via a linear transformation; (2) Across different networks, their most active and least active filters (shown in Figures S11 and S12 in Li et al., 2016) were quite different, indicating different neuron selectivity. In sum, the results in Li et al. (2016) indicate that by merely optimizing the image categorization performances, different HCNNs can obtain different object representations but with similar categorization performance. This seems not consistent with the goal-driven principle.

In Hong et al. (2016), the authors were aware of this divergent HCNN learning problem. They said, quote: "It is not the case that any deep convolutional network trained to solve an arbitrary object categorization task will trivially exhibit the features of ventral stream that are produced in our original high-variation-trained computational model." The authors seem to contribute such divergent-learning problems to the insufficiency of stimulus variations to stimulate IT neural sites. However, initialization is an inherent problem for HCNN learning, and it is not related to any external stimulus variations. In sum, HCNN architecture, initialization, learning algorithm, and training images all affect the outputs of the trained HCNN. We thought if different architectures are allowed, more divergent than convergent representations would be learnt, considering the existence of many local minima and the over-parameterization nature of HCNNs (LeCun et al., 2015). Hence "purely goal-driven" should be taken with great care in modeling sensory cortex.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB02070002), and National Natural Science Foundation of China (Grant Nos. 61421004, 61375042, 61573359, 61672489).

REFERENCES

Hong, H., Yamins, D. L., Majaj, N. J., and DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19, 613–622. doi: 10.1038/nn.4247

Khaligh-Razavi, S., and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Comput. Biol.* 10:e1003915. doi: 10.1371/journal.pcbi.1003915

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," *In Proceeding of Advances in Neural Information Processing Systems 25* (Lake Tahoe). 1106–1114.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. (2016). Convergent Learning: do different neural networks learn the same representations? arXiv:1511.07543v3.

Russakovsky, O., Deng, J., Su, H., Krause J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yamins, D. L., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi: 10.1038/nn.4244

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.