Journal Club

Editor's Note: These short, critical reviews of recent papers in the *Journal*, written exclusively by graduate students or postdoctoral fellows, are intended to summarize the important findings of the paper and provide additional insight and commentary. For more information on the format and purpose of the Journal Club, please see http://www.jneurosci.org/misc/ifa_features.shtml.

Encoding Voxels with Deep Learning

Panqu Wang,[1] Vicente Malave,[2] and Ben Cipollini[3]

Departments of [1]Electrical and Computer Engineering, [2]Cognitive Science, and [3]Computer Science and Engineering, University of California San Diego, La Jolla, California 92093

Review of Güçlü and van Gerven

Humans achieve fast and accurate recognition of complex objects through the ventral visual stream, a system of interconnected brain regions capable of hierarchical processing of increasingly complex features. In the feedforward view of the ventral visual stream, processing starts at the primary visual cortex V1, is carried through V2 and V4, and eventually reaches the inferior temporal (IT) cortex where more invariant and categorical representations useful for object identity are achieved (DiCarlo and Cox, 2007).

Past studies have shown that the receptive field of V1 complex cells are well characterized by 2D Gabor filters (Jones and Palmer, 1997; Carandini et al., 2005) and that IT contains subregions that are more activated for specific stimulus category, such as faces [fusiform face area (FFA) and occipital face area (OFA); Kanwisher et al., 1997], words [visual word form area (VWFA); McCandliss et al., 2003)], and scenes [parahippocampal place area (PPA); Epstein et al., 1999]. However, it remains unclear what intermediate features are represented in downstream areas between V1 and IT (Cox, 2014), how the representations can be quantified, and how representations at different points in the ventral visual stream interrelate.

Received Sept. 15, 2015; revised Oct. 24, 2015; accepted Oct. 30, 2015.

This work was partly funded by National Science Foundation (NSF) Grant SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center, and NSF grant IIS-1219252.

The authors declare no competing financial interests.

Correspondence should be addressed to Ben Cipollini, 9500 Gilman Dr, MC 0404, La Jolla, CA 92093. E-mail: bcipolli@ucsd.edu.

DOI:10.1523/JNEUROSCI.3454-15.2015

 0270-6474/15/3515769-03$15.00/0

Recently, researchers have used deep neural networks (DNNs) to probe representations in neural data, especially in IT (Cadieu et al., 2014; Yamins et al., 2014; Agrawal et al., 2014). DNNs are the state-of-the-art machine learning tool to solve computer vision tasks such as image classification (Krizhevsky et al., 2012; Szegedy et al., 2015), object detection (Girshick et al., 2014), and scene recognition (Zhou et al., 2014). These models stack computations, building "feature maps" by computing 2D filter responses at a number of input locations ("convolution") or taking the maximum response from a local subset of upstream inputs ("max-pooling"). With massive multilayer computations using millions of parameters to learn from millions of images, these systems can compete with human performance on these tasks (Cireşan et al., 2012; Taigman et al., 2014) and are inspired by the types of neural computations and feedforward component of the ventral visual stream.

Representations in the brain are also being probed with encoding and decoding models. Encoding and decoding models are complementary methods, with their relationship most easily understood in terms of the transforms they apply to an input space (e.g., stimuli), feature space (some abstract representation), and activity space [e.g., blood oxygenation level-dependent signal (BOLD) response; Naselaris et al., 2011]. Linearizing encoding models use a nonlinear transform to map from the input space (stimuli) to an abstract feature space (e.g., Gabor filters, as in Kay et al., 2008) that the brain is hypothesized to use, then use a linear transformation from the feature space to the activity space to see how well the brain's activity matches the encodings of the hypothesized abstract feature space. Linearizing decoding models also assume a nonlinear mapping between the input space (stimuli) and feature space (e.g., the categories of the input stimulus); instead of computing such a transform, they test whether the brain has done such a transform by finding the best linear mapping from brain activity to the feature space. The quality of the mapping between the activity space and feature space gives insight into how closely the brain's activity is underwritten by the features that define the feature space.

Until recently, most encoding models have used a predefined feature space or set of features. Güçlü and van Gerven (2014) and Agrawal et al. (2014) both learned feature spaces from a diverse set of complex naturalistic images, betting that the features underlying the brain's activity space reflect the image statistics of the world. Using unsupervised learning, Güçlü and van Gerven (2014) showed the learned features match the BOLD responses in early visual cortex better than the Gabor filters used by Kay et al. (2008). Agrawal and colleagues (2014) built encoding models from DNNs to investigate how well the DNN responses to naturalistic images could predict BOLD responses in the ventral visual stream.

Building from this recent work, Güçlü and van Gerven (2015) undertook a systematic examination of the predictability

for neural populations along the entire ventral visual stream. The authors used encoding models built on DNNs to predict BOLD responses in the ventral visual stream, then used multiple methods, including decoding models, to understand how the structure of the hierarchical neural network model mapped onto the ventral visual stream.

The encoding models were similar to that of Agrawal et al. (2014), who built encoding models from each layer of a multilayer convolutional neural network trained on millions of images (ImageNet dataset; Deng et al., 2009). Each encoding model was trained on 1750 images and tested on 120 images, and the best encoding model for each voxel was selected; the DNN layer of that encoding model is the voxel's optimal layer. Voxels that achieved test performance at or below chance were discarded. The decoding models estimated the most likely input stimulus by minimizing the distance between the measured brain response and the activity of the predicted brain response. The voxels were grouped by which neural network layer their activation patterns most closely matched.

These models showed three crucial results. First, they achieved excellent encoding and decoding performance. For the encoding model, the correlation (Pearson's r) between the observed and predicted response on the test set was 0.3∼0.51 for Subject 1 (S1) and 0.26∼0.42 for Subject 2 (S2), respectively. The decoding model achieved 98% and 93% accuracy (S1 and S2, respectively) on the test set, and surpassed the performance of previous models. Second, voxels followed a gradient in complexity: voxels in early visual areas (V1, V2) can be assigned to low-level features (edge, contrast), and voxels in downstream visual areas [lateral occipital complex (LO)] could be assigned to more complex features (object part, entire objects), as determined by human subject judgments of visualized neural network features (Zeiler and Fergus, 2014). DNNs trained on object recognition accounted for this gradient quantitatively, via estimates of layer complexity or assignment of voxels to the layer of their best encoding model. Finally, only voxels with higher-level features contributed to an object categorization task (animate vs inanimate), suggesting that object recognition is a guiding principle in the functional organization of the primate ventral stream.

The advances this paper makes in relating DNNs to neuroimaging data demonstrate how deep neural networks can advance computer vision, engineering more broadly, and our understanding of what representations the brain forms during tasks. Below, we outline three research directions this paper suggests.

Decomposing voxels into cortical computations

Encoding models in this paper, like those referenced above, draw correspondences between an abstract feature space and voxel activations but stop before drawing correspondences to meaningful computational units in the cortex. Voxel boundaries are an arbitrary byproduct of scanner settings and a subject's position in the scanner; they do not correspond to any meaningful functional unit of cortex. At the scale of the voxel ($2 \times 2 \times 2.5$ mm in this study), the most relevant computational unit in cortex is the macrocolumn (generally 0.5–1 mm in diameter in humans; Galuske et al., 2000). With the above voxel size, a voxel therefore represents approximately four to eight macrocolumns.

The paper by Güçlü and van Gerven (2015) pushes us even closer to drawing such specific correspondences. The encoding model computed a linear mapping of all features from a single network layer, meaning each voxel's activity is a superposition of feature maps—not too far from an explicit model of overlapping macrocolumn responses. If an encoding model attempted to explicitly model the hemodynamic interactions of macrocolumns or how features are represented in macrocolumns versus neurons, then an explicit correspondence between cortical computations and voxel activations could be tested. fMRI data at different resolutions (multi-millimeter, millimeter, submillimeter) could be used to test voxel responses built from varying degrees of population responses.

The ventral visual pathway is richer than suggested by comparison to a hierarchical DNN

Another opportunity suggested by the work of Güçlü and van Gerven (2015) lies behind the conclusions that the ventral visual stream is hierarchically organized. A more accurate claim may be that the ventral visual stream is capable of hierarchical processing and representation, but that the degree of hierarchy is task-dependent. Electrophysiological study shows dynamic interactions between areas, with multiple interacting parallel pathways that can operate even with complete damage to particular cortical areas in the so-called hierarchy (Kravitz et al., 2013). It is worth repeating that, in Kravitz et al. (2013), the task design targets feedforward processing. The stimulus is flashed briefly and unpredictably in both timing and content, and frequently followed by a backwards mask; these are used to reduce recurrent processing. Thus, it is not that the ventral visual stream is strictly hierarchical, but rather that these highly engineered tasks are most efficiently solved by the dynamic, interactive system in a fast, feedforward fashion.

DNNs are not limited to feedforward architectures, however, and Güçlü and van Gerven (2015) provide an analysis framework for doing similar studies using deep recurrent neural networks (DRNNs) on more ecologically valid tasks. DRNNs have had great success in natural language processing (Graves et al., 2013) and even generating image captions and descriptions (Socher et al., 2014; Fang et al., 2015). DRNNs could be trained on dynamic stimuli to examine the superior temporal sulcus, or movies to examine representations in time (Hasson et al., 2008). Such a program could expand our view on neural representations beyond the brief display of static images that currently dominates the literature. The paper by Güçlü and van Gerven (2015) outlines a framework capable of being extended to such paradigms.

Finally, the ventral visual stream is not a single pathway. In addition to its critical role in general object recognition, the ventral visual stream participates in fine-grained recognition (such as face identification) in OFA/FFA, word recognition in the VWFA, and scene recognition in the PPA. DNNs have been used to investigate many of these separately (Zhou et al., 2014; Yildirim et al., 2015), but a more complicated model may be needed to fully explain the data in ventral visual pathway. Furthermore, the ventral pathway may be reached through other pathways, such as a fast feedback system via frontal cortex (Bar, 2003). Clearly, much more can be done in this area.

How should we think of discarded voxels?

As mentioned above, Güçlü and van Gerven (2015) only used voxels which had a prediction accuracy significantly above chance by any encoding model in subsequent analyses. Consequently, >75% of voxels from V1, V2, V4, and LO, and >90% of voxels overall, were discarded from this study. Discarding this much of

the data deserves more explanation and could prevent the generalization of the findings to the entire ventral visual stream.

Notably, the distribution of discarded voxels may not be uniform. From the cortical maps (Güçlü and van Gerven, 2015, their Figs. 2*A*, 4*A*) it appears that the central visual field (CVF) is selectively underrepresented, as colored voxels are less present at the center of V1, V2, V4, and LO (central part of the map) than at the periphery. While the authors note that the majority of voxel receptive fields that show above-chance encoding performance have receptive fields at the center of the CVF (Güçlü and van Gerven, 2015, their Fig. 3*B*), this does not tell us what percentage of CVF voxels were discarded; the higher number of retained CVF voxels may simply be due to the cortical expansion of the CVF—there are more voxels with CVF receptive fields overall.

The authors suggest that discarded voxels are at least partially due to low signal to noise ratio (Güçlü and van Gerven, 2015, their Fig. 2*C*). However, the correlation between the two is relatively low ($r = 0.27$ for S1, 0.22 for S2), suggesting this explanation is incomplete. The reasons for the large numbers of discarded voxels should be explored further in future studies.

In conclusion, in decontextualized object recognition tasks, the computations and architecture of deep neural networks have correspondences to the human ventral visual stream. The use of encoding models and deep neural networks holds promise to provide deeper insight into relationships between neurons, voxels, and cortical representations of visual information.

References

Agrawal P, Stansbury D, Malik J, Gallant JL (2014) Pixels to voxels: modeling visual representation in the human brain. arXiv 1407.5104 [q-bio.NC].

Bar M (2003) A cortical mechanism for triggering top-down facilitation in visual object recognition. J Cogn Neurosci 15:600–609. CrossRef Medline

Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10:e1003963. CrossRef Medline

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC (2005) Do we know what the early visual system does? J Neurosci 25:10577–10597. CrossRef Medline

Cireşan D, Meier U, Masci J, Schmidhuber J (2012) Multi-column deep neural network for traffic sign classification. Neural Netw 32: 333–338. CrossRef Medline

Cox DD (2014) Do we understand high-level vision? Curr Opin Neurobiol 25:187–193. CrossRef Medline

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. IEEE Conf Comp Vis Pattern Recogn 2009:248–255. IEEE

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11: 333–341. CrossRef Medline

Epstein R, Harris A, Stanley D, Kanwisher N (1999) The parahippocampal place area: recognition, navigation, or encoding? Neuron 23:115–125. CrossRef Medline

Fang H, Gupta S, Iandola F, Srivastava R, Deng L, Dollar P, Gao J, He X, Mitchell M, Platt J, Zitnick L, Zweig G (2015) From captions to visual concepts and back. IEEE Conf Comp Vis Pattern Recogn 2015:1473–1482.

Galuske RA, Schlote W, Bratzke H, Singer W (2000) Interhemispheric asymmetries of the modular structure in human temporal cortex. Science 289:1946–1949. CrossRef Medline

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conf Comp Vis Pattern Recogn 2014:580–587.

Graves A, Mohamed AR, Hinton G (2013) Speech recognition with deep recurrent neural networks. IEEE Internat Conf Acoust Speech Signal Process 2013:6645–6649.

Güçlü U, van Gerven MA (2014) Unsupervised feature learning improves prediction of human brain activity in response to natural images. PLoS Comput Biol 10:e1003724. CrossRef Medline

Güçlü U, van Gerven MA (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci 35:10005–10014. CrossRef Medline

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. J Neurosci 28: 2539–2550. CrossRef Medline

Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58:1233–1258. Medline

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311. Medline

Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352–355. CrossRef Medline

Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M (2013) The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends Cogn Sci 17:26–49. CrossRef Medline

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neur Inform Process Systems NIPS Proceed 2012:1097–1105.

McCandliss BD, Cohen L, Dehaene S (2003) The visual word form area: expertise for reading in the fusiform gyrus. Trends Cogn Sci 7:293–299. CrossRef Medline

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400–410. CrossRef Medline

Socher R, Karpathy A, Le QV, Manning CD, Ng AY (2014) Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comp Linguis 2:207–218.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. IEEE Conf Comp Vis Pattern Recogn 2015:1–9.

Taigman Y, Yang M, Ranzato M, Wolf L (2014) DeepFace: closing the gap to human-level performance in face verification. IEEE Conf Comp Vis Pattern Recogn 2014:1701–1708.

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111:8619–8624. CrossRef Medline

Yildirim I, Kulkarni TD, Freiwald WA, Tenenbaum JB (2015) Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. Proceed Ann Conf Cogn Sci Soc 2015:2751–2756.

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. Eur Conf Comp Vis Proc 2014:818–833.

Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. NIPS Proceed 2014:487–495.