Eight open questions in the computational modeling of higher sensory cortex

Propelled by recent advances in biologically-inspired computer vision and artificial intelligence, the past five years have seen significant progress in using deep neural networks to model response patterns of neurons in higher visual cortical areas. In this paper, we briefly review this progress and then discuss eight key “open questions” that we believe will drive research in computational models of sensory systems over the next five years, both in visual cortex and beyond. Throughout, our focus is on challenges that will require both cutting-edge algorithmic developments as well as next-generation neuroscience and cognitive science experiments.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

Eight Open Questions in the Computational Modeling of Higher Sensory Cortex

Daniel L. K. Yamins[1] and James J. DiCarlo[1]

[1]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Propelled by advances in biologically-inspired computer vision and artificial intelligence, the past five years have seen significant progress in using deep neural networks to model response patterns of neurons in visual cortex. In this paper, we briefly review this progress and then discuss eight key "open questions" that we believe will drive research in computational models of sensory systems over the next five years, both in visual cortex and beyond.

Any scientific development of long-term value opens up as many new questions as it answers. This is certainly the case with recent progress in building deep neural network models of visual cortex. In this piece, our goal is to briefly describe these recent advances, and to outline what we consider to be the most interesting open problems in cortical modeling, both in vision and beyond. Throughout, our focus is on questions that will require both cutting-edge algorithmic developments as well as next-generation neuroscience and cognitive science experiments.

Brief Review of Recent Progress

Starting with the seminal ideas of Hubel and Wiesel, work in visual systems neuroscience over the past 60 years has shown that the ventral visual stream generates invariant object recognition behavior via a hierarchically-organized series of cortical areas that encode object properties with increasing selectivity and tolerance (*1–5*). Early visual areas, such as V1 cortex, capture low-level features such as edges and center-surround patterns (*6, 7*). In contrast, neural population responses in the highest ventral visual areas, inferior temporal (IT) cortex, can be used to decode object category, robust to significant variations present in natural images (*8–10*). The featural content of mid-level visual areas such as V2, V3, and V4 is less well understood, but these areas appear to contain intermediate computations between simple edges and complex objects, along a pipeline of increasing receptive field sizes (*1, 11–18*).

Many of these observations can be captured mathematically via class of computational architectures known as Hierarchical Convolutional Neural Networks (HCNNs), a generalization of Hubel and Wiesel's simple and complex cells that has been developed over the past 30 years (*19, 20*). HCNN models are composed of several retinotopic layers combined in series. Each layer is very simple, but together they produce a deep, complex transformation of the input

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

data — in theory, like the transformation produced in the ventral stream. However, mapping a single HCNN model to ventral stream neural data has proven extremely challenging (*12*), in part because subtle parameter changes (eg. number of layers, local receptive field sizes, &c) can dramatically affect a model's match to neural data (*21, 22*). Recent work in visual cortex seeks to go beyond this powerful but broad-stroke understanding to identify concrete predictive models of ventral cortex, and then use these models to gain insight inaccessible without large-scale computational precision.

A key aspect of this approach has been *performance-based* optimization, in which the parameters of a large multi-layer neural networks are chosen to optimize the networks' performance on a high-level, ecologically valid visual task (*23*). Leveraging computer vision and machine learning techniques, together with large amounts of real-world labelled images used as supervised training data, (*24–26*), HCNNs have been produced that achieve near-human-level performance on challenging object categorization tasks (*27*).

Intriguingly, even though these networks are not directly optimized to fit neural data, their top hidden layers are nonetheless highly predictive of single-site neural responses as well population-level representations in IT cortex both in electrophysiological (*23, 28*), and fMRI data (*29, 30*). Specifically, model units from the highest hidden layers of these performance-optimized HCNN can be linearly combined to produce synthetic "neurons" that predict the image-by-image response patterns of sites in IT cortex. Moreover, the population of these synthetic neurons closely matches the representational dissimilarity matrices (RDMs, (*31*)) of the macaque and human IT populations. These deep, performance-optimized neural networks have thus yielded the first quantitatively accurate, predictive model of the IT population response.

Moreover, high-throughput computational experiments evaluating thousands of HCNN models on both task performance and neural-predictivity metrics, have found a strong correlation between performance of high-level object recognition tasks and ability to explain IT cortical spiking data (*23*). The predictive power of these models is driven not just by categorization performance alone, as ideal observer models with perfect access to object identity do not themselves predict IT neural response patterns nearly as well as the hierarchical neural network units (*23*).

Critically, these HCNN models are *mappable* not only to IT, but also to other levels of the ventral visual stream. Lower model layer filter weights resemble Gabor wavelets and are effective models of fMRI voxel responses in V1 voxel data (*29, 30*). Along the same lines, intermediate HCNN layers are predictive of neural responses in V4 cortex (*23*). In other words: combining two general biological constraints — the behavioral constraint of recognition performance, and the architectural constraint imposed by the HCNN model class — leads to greatly improved models of multiple layers of the visual sensory cascade. An additional benefit of this approach is that each layer of the HCNN is a *basis set* for its corresponding cortical area, from which large numbers of IT-, V4- or V1-like units can be generated. A common assumption in visual neuroscience is that understanding the qualitative structure of tuning curves in lower cortical areas (e.g. gabor conjunctions in V2 or curvature in V4 (*32*)) is a necessary precursor to explaining higher visual cortex. These recent results show that higher-level constraints can yield quantitative models even when bottom-up primitives have not yet been identified.

The mapping between neural networks and cortical neural responses is still far from perfect. However, these recent results are encouraging, and they advance the understanding of the ventral stream in at least two new ways. First, the predictive accuracy of these models suggests that

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

the principles of cortical processing may be best described at the level of architectural statistics (rather than precise wiring patterns), learning rules (rather than descriptors of tuning curves), and ethological task goals (rather than information transmission). Second, because models derived from this approach are both accurately predictive and generative, they act as hypothesis generators that can be richly interrogated to explore key open questions and enable the rational design of neuroscience experiments to answer those questions. Below we list eight exciting open questions that are now approachable from this new vantage point.

1 Why is IT cortex heterogenous at large spatial scales?

IT cortex is not a single monolithic computational mass in which output features are randomly intermixed across the cortical surface, but instead is likely to contain multiple retinotopic areas, with posterior IT, central IT, and anterior IT areas performing potenially different computations (*1*). It is also now known that specialized face, place, body, and color-preferring regions at the multiple-millimeter scale are found in each of these IT areas (*33–36*). Are these the only regions? If so, why these and not others? How do the regions arise in the first place? Understanding this heterogeneity with computational models has two components: first, identifying whether and how the observed distributions of unit selectivities arise, independently of their spatial clustering; and second, explaining the observed spatial clustering.

Existing HCNN models could likely be used to generate detailed predictions about the unit distributions. A basic question is: to what extent are the existence of apparently specialized populations of units (e.g. face-selective units) strongly dependent on the semantic content of the training data of the neural networks? Will standard neural network model approaches yield observed unit populations if trained on datasets with a mix of semantic content close to that experienced by humans during development (e.g. a large fraction of faces)? How sensitive are unit selectivity distributions to this semantic content?

The second question, about spatial clustering, will require a more substantial extension of the HCNN framework, since those models make no specific predictions about how their units are to be mapped to the two-dimensional cortical sheet. It is possible that using a simple self-organizing map approach (*37*) to cluster in space units with similar feature tunings would explain a large fraction of the spatial structure in IT. However, there is some evidence that clustering may not be along purely geometric or featural lines — e.g., body-preferring patches arise near face-preferring patches even though there is no obvious geometric similarity between these two categories (*38*). If the known regions do not emerge in these types of models, it will be important to understand what additional principles are required to build them. If they do, it will also be of interest to search for new model-predicted regions that could subsequently be confirmed or falsified using primate fMRI and electrophysiolgy experiments.

2 Which visual properties are explicitly encoded in intermediate ventral stream areas?

Intermediate visual areas such as V2 and V4 have proven especially hard to understand because, unlike V1 and IT, they are removed both from low-level image properties and from higher-level human semantic intuition (*32, 39, 40*). Because intermediate layers of computational models are predictive of these cortical areas, performing high-throughput "virtual electrophysiology"

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

to characterize the model's internal structure should yield insight into tuning curves in corresponding cortical areas. One of the key advantages of an image-computable model is that it can be analyzed in great detail at low cost. It will be of great interest to perform high-throughput "virtual electrophysiology" on models to determine what their intermediate units represent. Recent progress in visualization techniques have allowed for structures in intermediate layers to be explored. Techniques that seek to optimize input images either for matching the statistics of existing images or for the activation of a unit (or units) within model higher layers have been seen to produce impressive results in texture generation, image style matching, and model exploration (*41–43*). Versions of these techniques could have useful application in neuroscience for efficiently separating multiple candidate models, by optimizing for stimuli that produce the largest difference between two such models, even if, and perhaps most especially if, those alternative models do not differ in simple conceptual ideas. Such techniques (inspired by (*9, 44*)) could be extremely helpful in reducing the huge stimulus space to a set that would be small enough to measure neural responses using existing experimental techniques.

3 Can models predict the perceptual consequences of direct neural perturbations?

Computational models enable the efficient exploration of the effects of perturbing unit activations in a highly selective manner. In the short term, such an approach could be used to make testable predictions of the behavioral changes (eg. in facial expression identification ability) that arise from inactivating/stimulating specific subpopulations of units defined by cell type or functionality (eg. high face-selectivity). Over the next 2 – 5 years, it should be possible to combine computational models with cutting-edge optical techniques in non-human primates (*45, 46*) to help design highly targeted real-time neural perturbation studies.

4 Is "linear readout" a real model of downstream neural decoding?

In the above work, it is often implicitly assumed that behavior is generated from neural representations by *linear readout*, e.g. neurons in other downstream brain areas (e.g. PFC, motor cortex, parietal cortex) using the information in the high-level representational areas (e.g. IT) by forming linear combinations of neurons from those areas. In other words, linear classifiers are not merely thought of as "information measurement" devices, but also as a concrete hypothesis for how multiple downstream brain areas could utilize the robust, explicit information in a high-level representation for multiple task purposes (*8, 47, 48*). Evaluating this hypothesis more thoroughly is important, involving a variety of questions, including:

- To what extent is the hypothesis true? Are downstream visual readouts really linear as (*8*) might suggest, or for some tasks do they need to be (at least somewhat) nonlinear? How important are temporal coding mechanisms in visual task readout?
- Building on (*47*), can we identify one or several examples of these linear readouts in action, connecting IT to some specific downstream area? What types neurons, in which cortical layers, are involved in the readout process?
- How are linear readouts learned, and over what timescales do they form? For them to play the role they are often assigned in the current models, such readouts would have to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

form comparatively quickly, with small amounts of on-line training data. Can the process of fast learning be captured empirically? What specific rules of learning are used? Are the classifier weight learning rules implicated more like those using in machine learning classifier algorithms, such as Support Vector Machines (SVMs), or some simpler procedure like linear discriminants (*49*)? If there really are linear classifier learning algorithms running neurally, do their weightings on input neurons get regularized as they might in machine learning algorithms (*50*) — that is, discouraged from having strong connections from too many upstream neurons, beyond the constraints imposed by biophysics alone? If so, what type of sparsity priors are imposed by the regularization used in the real brain?

- What are the *default readouts* that run automatically in a typically inattentive or passive state? How is *task-switching* accomplished, and how does this relate to attentional state? Presumably the task attention of the animal is deeply connected to switching between the "active" readout at any given time. How is this accomplished mechanistically? How would we even identify the neural signature of "active" readout?

5 How do feedforward discriminative networks connect to top-down inference in generative models?

The HCNN models of vision mentioned above are largely *discriminative*, in that they model the flow of information from bottom-to-top: information comes in to the system along sensory input arrays (e.g. the retina) and exits via outputs that indicate an attribute of the data relative to an (often discrete) discrimination task (e.g recognizing a given category of object in the image). However, the *generative* point of view is also relevant and a potentially critical next step in sensory models. In this view, there is an explicit process through which high-level labels about the world (e.g. object identities, relation locations and relationships) can be turned into predictions about the low-level (pixel) description consistent with those high-level labels. Generative models are important because they naturally support a number of key cognitive phenomena, such as inference, explaining away, and imagination — and would be a natural source of "data" for self-supervised learning procedures. They also offer the potential of explaining the well-known, but poorly understood role of cortico-cortical feedback connections in the brain. Figuring out how to combine bottom-up discriminative models with more top-down generative models, and connect them both to the neuroscience of visual perception in ambiguous or complex images (like the "Dallenbach Cow" (*51*)) would be a significant step forward.

6 How is visual learning actually implemented in the brain?

While recent work has begun to uncover how images are encoded in adult IT cortex, very little is known about how the IT representation arises in the first place. To what extent does visual learning during development shape high-level vision? Computational models — which, at heart, are really the product of learning rules in action — can help us think about these key questions in a new way.

Three important questions associated with this are:

- What aspects of the visual system are evolved, vs. developed, vs. learned? How do the learning rules active during post-natal development differ from those active in expertise

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

learning in adulthood? Mapping learning trajectories in models to developmental data will help make predictions about these important questions.

- What types of data are learning rules receiving? A significant limitation of many of the most effective existing learning rules for computational models — such as error back-propagation for training categorization in deep neural networks — is that they require large amounts of high-level semantically labeled training data. Over the next few years, an important area of research will be the discovery of semi- or un-supervised neural network learning algorithms that blend features of existing machine-learning techniques with constraints from neural data. An intriguing possibility is that heavily supervised category-label training could be replaced by optimization for properties (e.g. position, size, and pose) that can be more easily estimated from motion heuristics in natural video.

- How are learning rules implemented at the circuit level? Existing results suggest a number of cost-function objectives that may be involved in perceptual learning, including stimulus reconstruction goals, efficient-coding constraints (*52*), and top-down supervised task prediction errors (*53*). These rule need to be implemented as a (presumably recurrent) part of the neural network that can be then used for tasks during "runtime". By combining high-resolution neural techniques such as calcium imaging and axonal tracing, it may be possible to watch learning in action and use these data to constrain an understand of how the algorithmic rules underlying this learning are implemented in neural circuits.

7 Can sensory systems beyond vision benefit from performance-driven neural-network approaches?

Recent work in vision suggests a more general hypothesis about how to model sensory cortex: selecting biologically-plausible neural networks for high performance on an ecologically-relevant sensory task will yield a detailed model of the actual cortical areas that underlie that task. Since this idea has some traction in the ventral visual stream, it is natural to ask whether it also yields insight in other sensory domains.

Some initial recent work has found that HCNNs trained to solve challenging high-variation word recognition tasks are predictive of voxel patterns in auditory cortex (*54*). These models are also able to differentiate auditory areas, with lower model layers more predictive of inferior colliculus, intermediate layers more predictive of primary auditory cortex, and higher layers of speech and music-selective areas identified in recent imaging studies (*55*).

These results open up a variety of computational audition questions, including:

- **Detailed characterization of non-primary auditory cortex.** As with higher visual areas, models can be used to make detailed testable predictions about poorly-understood auditory cortex subregions, especially in relation to speech and natural sound representations (*55, 56*). This problem could be approached experimentally using both human fMRI/eCOG and non-human primate electrophysiology techniques.

- **What auditory tasks best explain cortical differentiation?** Given the evolutionary history of audition, non-speech tasks (eg. environmental sound differentiation) might be as important for driving auditory cortex structure as speech. This question could be explored by training networks on a variety of ecological auditory tasks.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

- **How do audition-optimized architectures compare to those optimized for vision?** Are there deep but hidden structural similarities between visual and auditory cortex that arise from underlying similarities in auditory and visual data? This fascinating question could be attacked both from a purely algorithmic point of view, as well as by comparing auditory neural data to ventral-stream data.

It will also be of interest to see if similar ideas can used in other sensory domains. One promising idea would be to model neural responses in rat barrel cortex through hierarchical networks optimized for somatosensory tasks. It is also an intriguing question of how and whether this approach has utility for understanding other evolutionarily older sensory systems, eg. olfaction — where a key challenges will lie in framing behavioral tasks and architecture classes in the first place.

8 How should working memory be included? Beyond Sense-At-A-Glance

Largely feedforward models like HCNNs cannot provide a full account of the dynamics of brain systems that store extensible state – most notably, any brain area that involves working memory, since the "dynamics" of a feedforward network will always converge to the same state independent of input. However, in studying behavior downstream of the sensory system, e.g., the neural implementation of visual working memory, having a working model of the visual system inputs to memory areas would be of use. Moreover, there is a growing body of literature connecting recurrent neural networks to neural phenomenon in attention, decision making and motor program generation (*57–59*). Results from reinforcement learning (*60*) have shown how powerful such techniques may be in the realm of AI. Mapping these to ideas in the neuroscience of the interface between ventral visual cortex and (e.g.) parietal cortex or the hippocampus will be of great interest (*61, 62*). Models that combine rich sensory input systems, as modeled by deep neural networks, with these recurrent networks, could provide a fruitful avenue for exploring more sophisticated cognitive behaviors beyond simple categorization or binary decision making, breaking out of the pure "representation" framework in which sensory models are often cast. This is especially interesting when there could be a complex loop between behavioral output and input stimulus, e.g. when modeling exploration of an agent over long time scales in a complex sensory environment (*63*).

Conclusion

Recent computational advances have transformed our ability to accurately model the neural responses of sensory systems, even at high levels of the sensory hierarchy that were previously deeply mysterious. However, there is much more exciting work to be done, and, if the recent past is a guide to the future, the next successes will require the continued convergence of techniques and workers in neuroscience, machine learning, computer science, and psychophysics.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

References

1. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–34 (2012).

2. Malach, R., Levy, I. & Hasson, U. The topography of high-order human object areas. *Trends in cognitive sciences* **6**, 176–184 (2002).

3. Felleman, D. & Van Essen, D. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* **1**, 1–47 (1991).

4. Rust, N. C. & DiCarlo, J. J. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *J Neurosci* **30**, 12978–95 (2010).

5. Connor, C. E., Brincat, S. L. & Pasupathy, A. Transformation of shape information in the ventral pathway. *Curr Opin Neurobiol* **17**, 140–7 (2007).

6. Carandini, M. *et al.* Do we know what the early visual system does? *J Neurosci* **25**, 10577–97 (2005).

7. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of physiology* **283**, 53–77 (1978).

8. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *The Journal of Neuroscience* **35**, 13402–13418 (2015).

9. Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z. & Connor, C. E. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat Neurosci* (2008).

10. Hung, C. P., Kreiman, G., Poggio, T. & Dicarlo, J. J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).

11. Freeman, J. & Simoncelli, E. Metamers of the ventral stream. *Nature Neuroscience* **14**, 1195–1201 (2011).

12. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn Sci* **11**, 333–41 (2007).

13. Schmolesky, M. T. *et al.* Signal timing across the macaque visual system. *J Neurophysiol* **79**, 3272–8 (1998).

14. Lennie, P. & Movshon, J. A. Coding of color and form in the geniculostriate visual pathway (invited review). *J Opt Soc Am A Opt Image Sci Vis* **22**, 2013–33 (2005).

15. Schiller, P. Effect of lesion in visual cortical area v4 on the recognition of transformed objects. *Nature* **376**, 342–344 (1995).

16. Gallant, J., Connor, C., Rakshit, S., Lewis, J. & Van Essen, D. Neural responses to polar, hyperbolic, and cartesian gratings in area v4 of the macaque monkey. *Journal of Neurophysiology* **76**, 2718–2739 (1996).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

17. Brincat, S. L. & Connor, C. E. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat Neurosci* **7**, 880–6 (2004).

18. Yau, J. M., Pasupathy, A., Brincat, S. L. & Connor, C. E. Curvature processing dynamics in macaque area v4. *Cerebral Cortex* bhs004 (2012).

19. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybernetics* (1980).

20. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 255–258 (1995).

21. Pinto, N., Cox, D. D. & Dicarlo, J. J. Why is real-world visual object recognition hard? *PLoS Computational Biology* (2008).

22. Yamins, D., Hong, H., Cadieu, C. & Dicarlo, J. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. *Advances in Neural Information Processing Systems* (2013).

23. Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619–8624 (2014).

24. Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* (2012).

25. Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of The 30th International Conference on Machine Learning*, 115–123 (2013).

26. Deng, J., Li, K., Do, M., Su, H. & Fei-Fei, L. Construction and analysis of a large scale image ontology. In *Vision Sciences Society* (2009).

27. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

28. Cadieu, C. F. *et al.* Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology* **10**, e1003963 (2014).

29. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Comp. Bio.* (2014).

30. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience* **35**, 10005–10014 (2015).

31. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–41 (2008).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

32. Sharpee, T. O., Kouh, M. & Reyholds, J. H. Trade-off between curvature tuning and position invariance in visual area v4. *PNAS* **110**, 11618–11623 (2012).

33. Downing, P. E., Chan, A., Peelen, M., Dodds, C. & Kanwisher, N. Domain specificity in visual cortex. *Cerebral Cortex* **16**, 1453–1461 (2006).

34. Freiwald, W. A. & Tsao, D. Y. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–51 (2010).

35. Vaziri, S., Carlson, E. T., Wang, Z. & Connor, C. E. A channel for 3d environmental shape in anterior inferotemporal cortex. *Neuron* **84**, 55 – 62 (2014). URL http://www.sciencedirect.com/science/article/pii/S0896627314007442.

36. Lafer-Sousa, R. & Conway, B. R. Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nat Neurosci* **16**, 1870–1878 (2013). URL http://dx.doi.org/10.1038/nn.3555.

37. Kohonen, T. The self-organizing map. *Neurocomputing* **21**, 1–6 (1998).

38. Schwarzlose, R. F., Baker, C. I. & Kanwisher, N. Separate face and body selectivity on the fusiform gyrus. *The Journal of Neuroscience* **25**, 11055–11059 (2005).

39. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fmri. *Neuroimage* **56**, 400–410 (2011).

40. Nishimoto, S. & Gallant, J. L. A three-dimensional spatiotemporal receptive field model explains responses of area mt neurons to naturalistic movies. *The Journal of Neuroscience* **31**, 14551–14564 (2011).

41. Gatys, L. A., Ecker, A. S. & Bethge, M. A neural algorithm of artistic style (2015). URL http://arxiv.org/abs/1508.06576.

42. Gatys, L. A., Ecker, A. S. & Bethge, M. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. In *Advances in Neural Information Processing Systems 28* (2015). URL http://arxiv.org/abs/1505.07376.

43. Mordvintsev, A., Tyka, M. & Olah, C. Inceptionism: Going deeper into neural networks, google research blog (2015).

44. Paninski, L., Pillow, J. & Lewi, J. Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in brain research* **165**, 493–507 (2007).

45. Afraz, S. R., Kiani, R. & Esteky, H. Microstimulation of inferotemporal cortex influences face categorization. *Nature* **442**, 692–5 (2006).

46. Afraz, A., Boyden, E. S. & DiCarlo, J. J. Optogenetic and pharmacological suppression of spatial clusters of "face neurons" reveal their causal role in face discrimination. *(Under review)* (2014).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

47. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316 (2001).

48. Pagan, M., Urban, L. S., Wohl, M. P. & Rust, N. C. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nature neuroscience* **16**, 1132–1139 (2013).

49. Misaki, M., Kim, Y., Bandettini, P. A. & Kriegeskorte, N. Comparison of multivariate classifiers and response normalizations for pattern-information fmri. *Neuroimage* **53**, 103–118 (2010).

50. Evgeniou, T., Pontil, M. & Poggio, T. Regularization networks and support vector machines. *Advances in computational mathematics* **13**, 1–50 (2000).

51. Dallenbach, K. M. A puzzle-picture with a new principle of concealment. *The American journal of psychology* (1951).

52. Barlow, H. B. Possible principles underlying the transformations of sensory messages (1961).

53. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, 177–186 (Springer, 2010).

54. Kell, A., Yamins, D., Norman-Haignere, S. & McDermott, J. Functional organization of auditory cortex revealed by neural networks optimized for auditory tasks. In *Society for Neuroscience* (2015).

55. Norman-Haignere, S., Kanwisher, N. & McDermott, J. H. Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *The Journal of Neuroscience* **33**, 19451–19469 (2013).

56. Leaver, A. M. & Rauschecker, J. P. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *The Journal of Neuroscience* **30**, 7604–7612 (2010).

57. Chikkerur, S., Serre, T., Tan, C. & Poggio, T. What and where: A bayesian inference theory of attention. *Vision research* **50**, 2233–2247 (2010).

58. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

59. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience* **18**, 1025–1033 (2015).

60. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

61. Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016

62. Hulbert, J. & Norman, K. Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex* bhu284 (2014).

63. Stadie, B. C., Levine, S. & Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814* (2015).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1881v1 | CC-BY 4.0 Open Access | rec: 20 Mar 2016, publ: 20 Mar 2016