nature neuroscience

Using goal-driven deep learning models to understand sensory cortex

Daniel L K Yamins[1,2] & James J DiCarlo[1,2]

Fueled by innovation in the computer vision and artificial intelligence communities, recent developments in computational neuroscience have used goal-driven hierarchical convolutional neural networks (HCNNs) to make strides in modeling neural single-unit and population responses in higher visual cortical areas. In this Perspective, we review the recent progress in a broader modeling context and describe some of the key technical innovations that have supported it. We then outline how the goal-driven HCNN approach can be used to delve even more deeply into understanding the development and organization of sensory cortical processing.

What should one expect of a model of sensory cortex?

Brains actively reformat incoming sensory data to better serve their host organism's behavioral needs (**Fig. 1a**). In human vision, retinal input is converted into rich object-centric scenes; in human audition, sound waves become words and sentences. The core problem is that the natural axes of sensory input space (for example, photoreceptor or hair cell potentials) are not well-aligned with the axes along which high-level behaviorally relevant constructs vary. For example, in visual data, object translation, rotation, motion in depth, deformation, lighting changes and so forth cause complex nonlinear changes in the original input space (the retina). Conversely, images of two objects that are ecologically quite distinct—for example, different individuals' faces—can be very close together in pixel space. Behaviorally relevant dimensions are thus 'entangled' in this input space, and brains must accomplish the untangling[1,2].

Two foundational empirical observations about cortical sensory systems are that they consist of a series of anatomically distinguishable but connected areas[3,4] (**Fig. 1b**) and that the initial wave of neural activity during the first 100 ms after a stimulus change unfolds as a cascade along that series of areas[2]. Each individual stage of the cascade performs very simple neural operations such as weighted linear sums of inputs or nonlinearities such as activation thresholds and competitive normalization[5]. However, complex nonlinear transformations can arise from simple stages applied in series[6]. Since the original input entanglement was highly nonlinear, the untangling process must also be highly nonlinear.

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to D.L.K.Y. (yamins@mit.edu).

Received 26 October 2015; accepted 13 January 2016; published online 23 February 2016; doi:10.1038/nn.4244

The space of possible nonlinear transformations that the brains neural networks could potentially compute is vast. A major challenge in understanding sensory systems is thus systems identification: identifying which transformations the true biological circuits are using. While identifying summaries of neural transfer functions (for example, receptive field characterization) can be useful[7], solving this systems identification problem ultimately involves producing an encoding model: an algorithm that accepts arbitrary stimulus inputs (for example, any pixel map) and outputs a correct prediction of neural responses to that stimulus. Models cannot be limited just to explaining a narrow phenomenon identified on carefully chosen neurons, defined only for highly controlled and simplified stimuli[8,9]. Operating on arbitrary stimuli and quantitatively predicting the responses of all neurons in an area are two core criteria that any model of a sensory area must meet (see **Box 1**).

Moreover, a comprehensive encoding model must not merely predict the stimulus-response relationship of neurons in one final area, such as (in vision) anterior inferior temporal cortex. Instead, the model must also be mappable: having identifiable components corresponding to intermediate cortical areas (for example, V1, V2, V4) and, ultimately, subcortical circuits as well. The model's responses in each component area should correctly predict neural response patterns within the corresponding brain area (**Fig. 1c**).

Hierarchical convolutional neural networks

Starting with the seminal work of Hubel and Wiesel[10], work in visual systems neuroscience has shown that the brain generates invariant object recognition behavior via a hierarchically organized series of cortical areas, the ventral visual stream[2]. A number of workers have built biologically inspired neural networks generalizing Hubel and Wiesel's ideas (for example, refs. 11–15). Over time, it was realized that these models were examples of a more general class of computational architectures known as HCNNs[16]. HCNNs are stacks of layers containing simple neural circuit motifs repeated across the sensory input; these layers are then composed in series. (Here, "layer" is used in the neural network sense, not in the cortical anatomy sense.) Each layer is simple, but a deep network composed of such layers computes a complex transformation of the input data—analogous to the transformation produced in the ventral stream.

The motifs in a single HCNN layer

The specific operations comprising a single HCNN layer were inspired by the ubiquitously observed linear-nonlinear (LN) neural motif[5]. These operations (**Fig. 1c**) include (i) filtering, a linear operation that takes the dot product of local patches in the input stimulus with a set of templates, (ii) activation, a pointwise nonlinearity—typically either

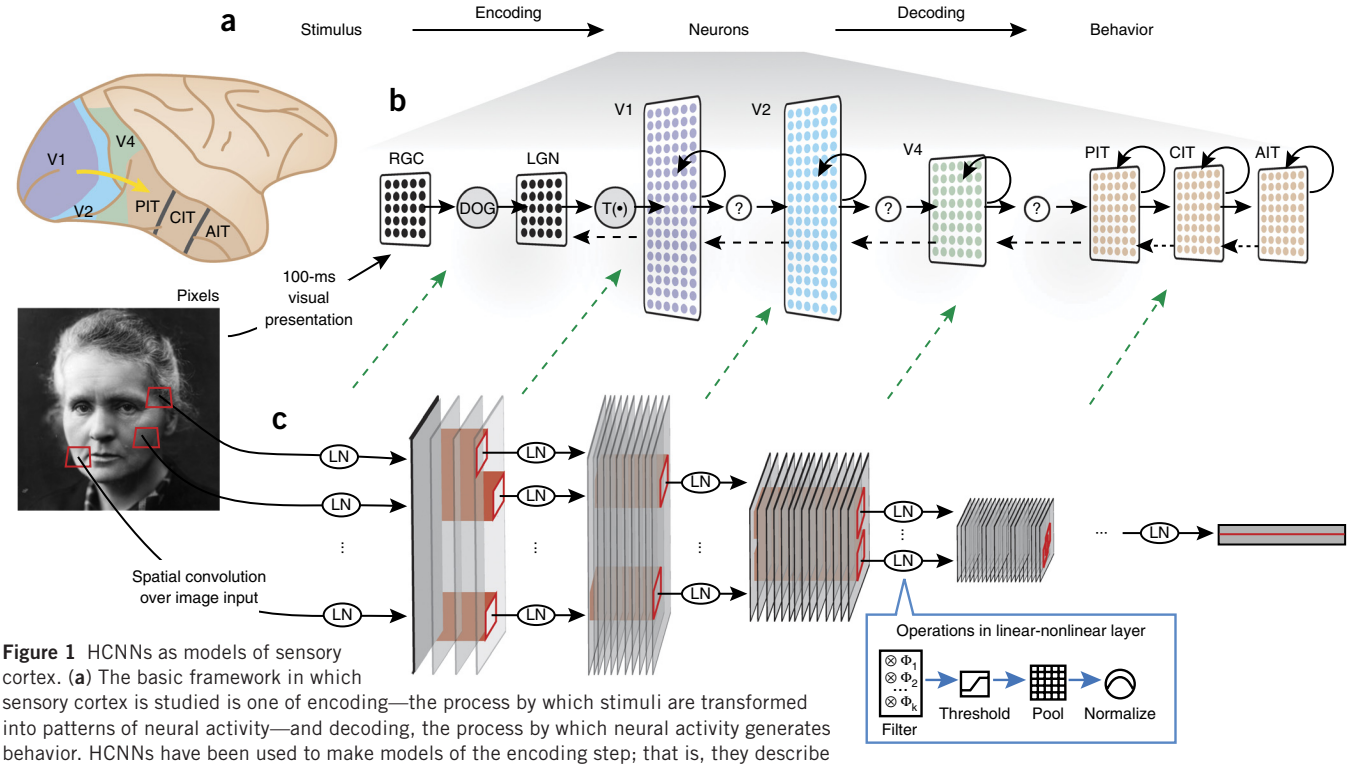

Figure 1 HCNNs as models of sensory cortex. (**a**) The basic framework in which sensory cortex is studied is one of encoding—the process by which stimuli are transformed into patterns of neural activity—and decoding, the process by which neural activity generates behavior. HCNNs have been used to make models of the encoding step; that is, they describe the mapping of stimuli to neural responses as measured in brain. (**b**) The ventral visual pathway is the most comprehensively studied sensory cascade. It consists of a series of connected cortical brain areas (macaque brain shown). PIT, posterior inferior temporal cortex; CIT, central; AIT, anterior; RGC, retinal ganglion cell; LGN, lateral geniculate nucleus. DoG, difference of Gaussians model; T(•), transformation. (**c**) HCNNs are multilayer neural networks, each of whose layers are made up of a linear-nonlinear (LN) combination of simple operations such as filtering, thresholding, pooling and normalization. The filter bank in each layer consists of a set of weights analogous to synaptic strengths. Each filter in the filter bank corresponds to a distinct template, analogous to Gabor wavelets with different frequencies and orientations; the image shows a model with four filters in layer 1, eight in layer 2, and so on. The operations within a layer are applied locally to spatial patches within the input, corresponding to simple, limited-size receptive fields (red boxes). The composition of multiple layers leads to a complex nonlinear transform of the original input stimulus. At each layer, retinopy decreases and effective receptive field size increases. HCNNs are good candidates for models of the ventral visual pathway. By definition, they are image computable, meaning that they generate responses for arbitrary input images; they are also mappable, meaning that they can be naturally identified in a component-wise fashion with observable structures in the ventral pathway; and, when their parameters are chosen correctly, they are predictive, meaning that layers within the network describe the neural response patterns to large classes of stimuli outside the domain on which the models were built.

a rectified linear threshold or a sigmoid, (iii) pooling, a nonlinear aggregation operation—typically the mean or maximum of local values[13], and (iv) divisive normalization, correcting output values to a standard range[17]. Not all HCNN incarnations use these operations in this order, but most are reasonably similar. All the basic operations exist within a single HCNN layer, which is then typically mapped to a single cortical area.

Analogously to neural receptive fields, all HCNN operations are applied locally, over a fixed-size input zone that is typically smaller than the full spatial extent of the input (**Fig. 1c**). For example, on a 256 × 256 pixel image, a layer's receptive fields might be 7 × 7 pixels. Because they are spatially overlapping, the filter and pooling operations are typically 'strided', meaning that output is retained for only a fraction of positions along each spatial dimension: a stride of 2 in image convolution will skip every second row and column.

In HCNNs, filtering is implemented via convolutional weight sharing, meaning that the same filter templates are applied at all spatial locations. Since identical operations are applied everywhere, spatial variation in the output arises entirely from spatial variation in the input stimulus. It is unlikely the brain literally implements weight sharing, since the physiology of the ventral stream and other sensory cortices appears to rule out the existence of a single master location in

Box 1 Minimal criteria for a sensory encoding model

We identify three criteria that any encoding model of a sensory cortical system should meet:
Stimulus-computability: The model should accept arbitrary stimuli within the general stimulus domain of interest;
Mappability: The components of the model should correspond to experimentally definable components of the neural system; and
Predictivity: The units of the model should provide detailed predictions of stimulus-by-stimulus responses, for arbitrarily chosen neurons in each mapped area.
These criteria may sometimes be in tension—insisting on mappability at the finest grain might hinder identifying models that actually work for complex real-world stimuli, since low-level circuit tools may operate best in reduced stimulus regimes. While seeking detailed models of neural circuit connectivity in simplified contexts is important, if such models do not add up in the aggregate to accurate predictors of neural responses to real-world stimuli, the utility of their lower-level verisimilitude is limited.

Box 2 Mapping models to neural sensory systems

How does one map artificial neural networks to real neurons? Several approaches are possible, at varying levels of neural detail.

Task information consistency. At the coarsest level, a useful metric of model similarity to a system is the consistency of patterns of explicitly decodable information available to support potential behavioral tasks. In this approach, populations of 'neurons' from a model and populations of recorded neurons are analyzed with identical decoding methods on a battery of high-level tasks (for example, object recognition, face identification and so forth). While not required, it is useful to use simple decoders such as linear classifiers or linear regressors[1,32,63,64], as these embody hypothetical downstream decoding circuits[65,66]. This procedure generates a pattern of response choices for both the model and the neural population. These patterns are then compared to each other either at a coarse grain (for example, via accuracy levels for each task[32]) or a fine grain (stimulus-by-stimulus response consistency). We note that this approach naturally connects to the linkage between neuronal populations and behavior[32], as both models and neurons can be compared to behavioral measurements from either in animal or humans subjects. Both the neural area thought be most directly connected to behavior (for example, IT in the visual case) and the computational model of this area should exhibit high consistency with those behavioral patterns[32].

Population representational similarity. Another population-level metric is representational similarity analysis[29,35], in which the two representations (that of the real neurons and that of the model) are characterized by their pairwise stimulus correlation matrix (**Fig. 2d**). For a given set of stimuli, this matrix describes how far apart a representation 'thinks' each pair of stimuli are. These distance matrices are then compared for similarity: the model is judged to be similar to the neural representation if it treats stimuli pairs as close to (or far from) each other whenever the real neural population representation also does so.

Single-unit response predictivity. A finer grained mapping of models to neurons is that of linear neural response predictivity of single units[33]. This idea is best understood via a simple thought experiment: imagine one had measurements from all neurons in a given brain area in two animals: a source animal and a target animal. How would one map the neurons in the source to neurons in the target? In many brain areas (such as, for example, V4 or IT), there might not be an exact one-to-one mapping of units between the animals. However, it is reasonable to suppose that the two animals' areas are the same (or very similar) up to linear transform—for example, that units in the target animal are approximately linear combinations of (a small number of) units in the source animal. In engineering terms, the animals would be said to be 'equivalent bases' for sensory representation. (If the mapping had to be nonlinear, it would call into question whether the two areas were the same across animals to begin with.) Making the mapping would, in effect, be the problem of identifying the correct linear combinations. The same idea can be used to map units in a model layer to neurons in a brain area. Specifically, each empirically measured neuron is treated as the target of linear regression from units in the model layer. The goal is find linear combinations of model units that together produce a 'synthetic neuron' that will reliably have the same response patterns as the original target real neuron: find $c_i, i \in \{1,\ldots,n\}$ such that

$$r(x) \approx r_{\text{synth}}(x) = \sum_i c_i m_i(x)$$

where $r(x)$ is the response of neuron r to stimulus x, and $m_i(x)$ is the response of the i-th model unit (in some fixed model layer). Accuracy of r_{synth} is then measured as its explained variance (R^2) for r on new stimuli not used to identify the coefficients c_i. Ideally, the number of model source units i that have nonzero weights c_i would be approximately the same as would be found empirically when attempting to map the neurons in one animal to those in same brain area for a different animal.

which shared templates could be stored. However, the natural visual (or auditory) statistics of the world are themselves largely shift invariant in space (or time), so experience-based learning processes in the brain should tend to cause weights at different spatial (or temporal) locations to converge. Shared weights are therefore likely to be a reasonable approximation to the brain's visual system, at least within the central visual field. The real visual system has a strong foveal bias, and more realistic treatment of nonuniform receptive field density might improve models' fits to neural data.

Deep networks through stacking

Since convolutional layer outputs have the same spatial layout as their inputs, output of one layer can be input to another. HCNNs can thus be stacked into deep networks (**Fig. 1c**). Although the local fields seen by units in a single layer have a fixed, small size, the effective receptive field size relative to the original input increases with succeeding layers. Because of repeated striding, deep HCNNs typically become less retinotopic with each succeeding layer, consistent with empirical observations[4]. However, the number of filter templates used in each layer typically increases. Thus, the dimensionality changes through the layers from wide and shallow to deep and narrow (**Fig. 1c**). After many strided layers, the spatial component of the output is so reduced that convolution is no longer meaningful, whereupon networks are typically extended using one or more fully connected layers. The last layer is usually used for readout: for example, for each of several visual categories, the likelihood of the input image containing an object of the given category might be represented by one output unit.

HCNNs as a parameterized model family

HCNNs are not a single model, but rather a parameterized model class. Any given HCNN is characterized by the following:

- discrete architectural parameters, including the number of layers the network contains, as well as, for each layer, discrete parameters specifying the number of filter templates; the local radius of each filtering, pooling and normalization operation; the pooling type; and potentially other choices required by the specific HCNN implementation; and
- continuous filter parameters, specifying the filter weights of convolutional and fully connected layers.

Though parameter choices might seem like mere details, subtle parameter differences can dramatically affect a network's performance on recognition tasks and its match to neural data[15,18].

Given the minimal model criteria described in **Box 1**, a key goal is identifying a single HCNN parameter setting whose layers correspond to distinct regions within the cortical system of interest (for example, different areas in the ventral stream) and which accurately predict response patterns in those areas (see **Box 2**).

While an oversimplification, the relationship between modifying filters and architectural parameters is somewhat analogous to that between developmental and evolutionary variation. Filter parameters are thought of as corresponding to synaptic weights, and their learning algorithms (see discussion of backpropagation below) update parameters in an online fashion. Changing architectural parameters, in contrast, restructures the computational primitives,

the number of sensory areas (model layers) and the number of neurons in each area.

Early models of visual cortex in context

A number of approaches have been taken to identify HCNN parameters that best match biological systems.

Hand-designing parameters via Hubel and Wiesel theory. Beginning in the 1970s, before the HCNN concept was fully articulated, modelers started tackling lower cortical areas such as V1, where neurons might be explicable through comparatively shallow networks. Hubel and Wiesel's empirical observations suggested that neurons in V1 resemble Gabor wavelet filters, with different neurons corresponding to edges of different frequencies and orientations[10,19]. Indeed, early computational models using hand-designed Gabor filter banks as convolution weights achieved some success in explaining V1 neural responses[20]. Later it was realized that models could be substantially improved using nonlinearities such as thresholding, normalization and gain control[5,21], helping motivate the HCNN class in the first place. Similar ideas have been proposed for modeling primary auditory cortex[22].

Learning parameters via efficient coding constraints. The work of Barlow, Olshausen and others introduced another way of determining filter parameters[23,24]. Filters were optimized to minimize the number of units activated by any given stimulus while still retaining the ability to reconstruct the original input. Such 'sparse' efficient codings naturally learn Gabor-wavelet-like filters from natural image data, without having to build those patterns in by hand.

Fitting networks to neural data. Another natural approach begun in the mid-1990s was to bring neuroscience data directly to bear on model parameter choice. The idea was to collect response data to various stimuli for neurons in a brain area of interest and then use statistical fitting techniques to find model parameters that reproduce the observed stimulus–response relationship. This strategy had some success fitting shallow networks to visual area V1, auditory area A1 and somatosensory area S1 (reviewed in ref. 25).

Difficulties with deeper networks. Given successful shallow convolutional models of early cortical areas, perhaps deeper models would shed light on downstream sensory areas. However, the deeper models needed to model such higher areas would have many more parameters than V1-like models. How should these parameters be chosen?

The outputs on which higher layers operate are challenging to visualize, making it difficult to generalize the hand-designed approach to deeper networks. Similarly, while some progress has been made in extending efficient coding beyond one layer[26], these approaches also have not yielded effective deeper networks. Multi-layer HMAX networks were created by choosing parameters roughly to match known biological constraints[12,13]. HMAX networks had some success reproducing high-level empirical observations, such as the tolerance ranges of inferior temporal (IT) cortex neurons[12,27] and the tradeoff between single-unit selectivity and tolerance[28].

However, by the mid-2000s, it had become clear that these approaches were all having trouble extending to higher cortical areas such as V4 and IT. For example, HMAX models failed to match patterns of IT population activity on batteries of visual images[29], while multilayered neural networks fit to neural data in V4 and IT ended up overfitting the training data and predicting comparatively small amounts of explained variance on novel testing images[8].

One plausible reason for this lack of success was that the largely feedforward neural networks being explored were too limited to capture the data efficiently. Perhaps more sophisticated network architectures, using feedback[30] or millisecond-scale spike timing[31], would be required. A second possibility was that failure arose from not having enough neural data to fit the model parameters. Single-unit physiology approaches[8] or whole-brain functional MRI[29] could measure responses to perhaps 1,000 independent stimuli, while array electrophysiology[32] could obtain responses to ~10,000 stimuli. In hindsight, the amount of neural data available to constrain such networks was several orders of magnitude too little.

A new way forward: goal-driven networks as neural models

The goal-driven approach is inspired by the idea that, whatever parameters are used, a neural network will have to be effective at solving the behavioral tasks the sensory system supports to be a correct model of a given sensory system. The idea of this approach is to first optimize network parameters for performance on an ethologically relevant task, and then, once network parameters have been fixed, to compare networks to neural data. This approach avoids the severe data limitation of pure neural fitting, as collecting (for example) millions of human-labeled images containing many hard real-world cases of object recognition is far easier than obtaining comparable neural data. The key question becomes: do such top-down goals strongly constrain biological structure? Will performance optimization imposed at the outputs of a network be sufficient to cause hidden layers in the network to behave like real neurons in, for example, V1, V4 or IT? A series of recent results has shown that this might indeed be the case.

The technological bases of the goal-driven approach are recent improvements in optimizing neural networks performance for artificial intelligence tasks. In this section, we discuss how these tools have led to better neural models; in the next, we discuss the technical innovations underlying those tools.

Top hidden layers of categorization-optimized HCNNs predict IT neuronal responses. High-throughput computational experiments evaluating thousands of HCNN models on task performance and neural-predictivity metrics revealed a key correlation: architectures that perform better on high-level object recognition tasks also better predict cortical spiking data[33,34] (**Fig. 2a**). Pushing this idea further by using recent advances from machine learning led to the discovery of hierarchical neural network models that achieved near-human-level performance level on challenging object categorization tasks. It turned out that the top hidden layers of these models were the first quantitatively accurate image-computable model of spiking responses in IT cortex, the highest-level area in the ventral hierarchy[18,33,34] (**Fig. 2b,c**). Similar models have also been shown to predict population aggregate responses in functional MRI data from human IT (**Fig. 2d**)[35,36].

These results are not trivially explained merely by any signal reflecting object category identity being able to predict IT responses. In fact, at the single neuron level, IT neural responses are largely not categorical, and ideal-observer models with perfect access to category and identity information are far less accurate IT models than goal-driven HCNNs[33] (**Fig. 2a,c**). Being a true image-computable neural network model appears critical for obtaining high levels of neural predictivity. In other words: combining two general biological constraints—the behavioral constraint of the object recognition task and the architectural constraint imposed by the HCNN model class—leads to greatly improved models of multiple layers of the visual sensory cascade.

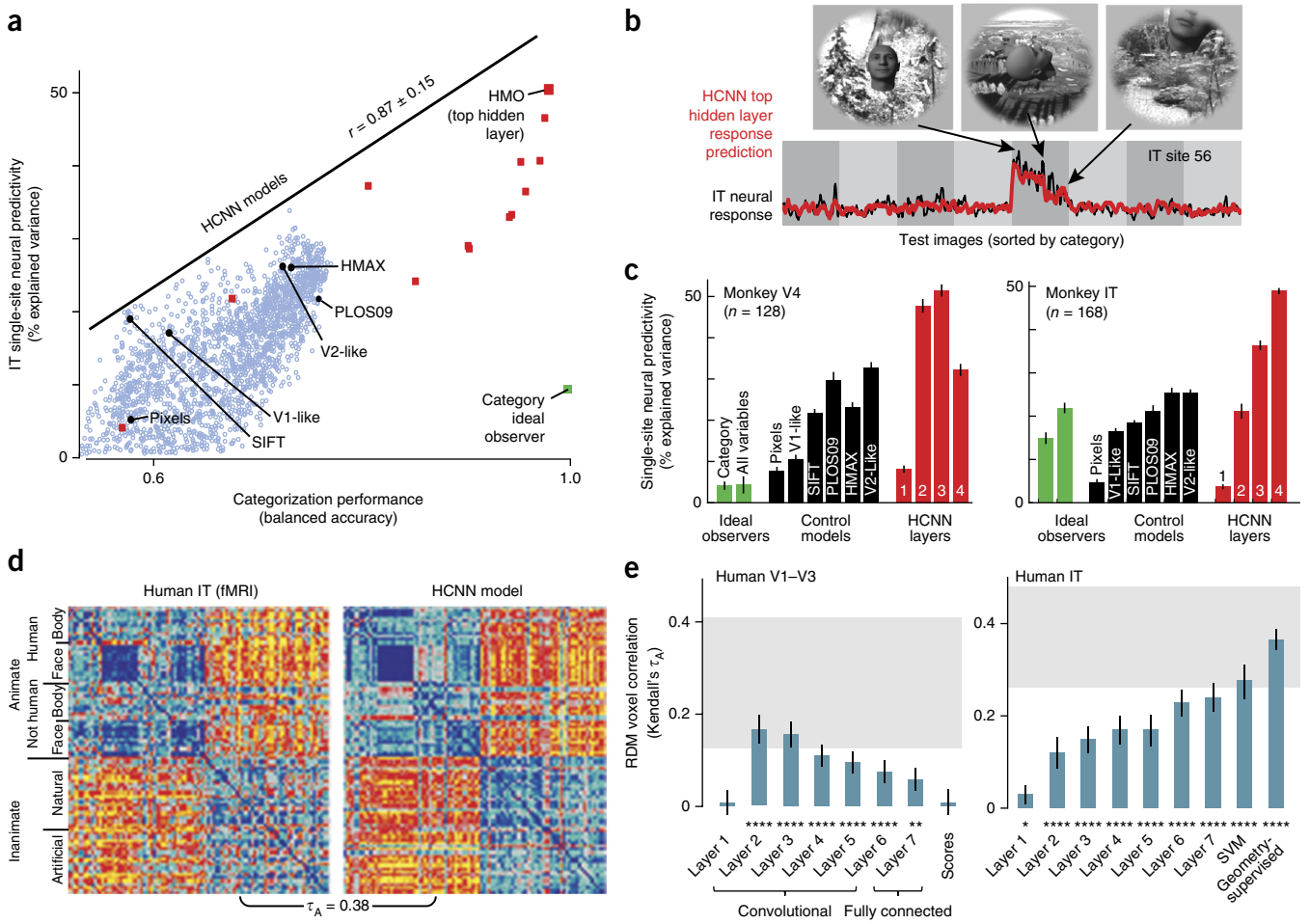

Figure 2 Goal-driven optimization yields neurally predictive models of ventral visual cortex. (**a**) HCNN models that are better optimized to solve object categorization produce hidden layer representations that are better able to predict IT neural response variance. The *x* axis shows performance (balanced accuracy; chance is 50%) of the model output features on a high-variation object categorization task. The *y* axis shows the median single-site IT response predictivity of the last hidden layer of the HCNN model, over $n = 168$ IT sites. Site responses are defined as the mean firing rate 70–170 ms after image onset. Response predictivity is defined as in **Box 2**. Each dot corresponds to a distinct HCNN model from a large family of such models. Models shown as blue circles were selected by random draws from object categorization performance-optimization; black circles show controls and earlier published HCNN models; red squares show the development over time of HCNN models produced during an optimization procedure that produces a specific HCNN model[33]. PLOS09, ref. 15; SIFT, shape-invariant feature transform; HMO, optimized HCNN. (**b**) Actual neural response (black trace) versus model predictions of the last hidden layer of an HCNN model (red trace) for a single IT neural site. The *x* axis shows 1,600 test images, none of which were used to fit the model. Images are sorted first by category identity and then by variation amount, with more drastic image transformations toward the right within each category block. The *y* axis represents the response of the neural site and model prediction for each test image. This site demonstrated face selectivity in its responses (see inset images), but predictivity results were similar for other IT sites[33]. (**c**) Comparison of IT and V4 single-site neural response predictivity for various models. Bar height shows median predictivity, taken over 128 predicted units in V4 (left panel) or 168 units in IT (right panel). The last hidden layer of the HCNN model best predicts IT responses, while the second-to-last hidden layer best predicts V4 responses. (**d**) Representational dissimilarity matrices (RDMs) for human IT and HCNN model. Blue color indicates low values, where representation treats image pairs as similar; red color indicates high values, where representation treats image pairs as distinct. Values range from 0 to 1. (**e**) RDM similarity, measured with Kendall's τ_A, between HCNN model layer features and human V1–V3 (left) or human IT (right). Gray horizontal bar represents range of performance of the true model given noise and intersubject variation. Error bars are s.e.m. estimated by bootstrap resampling of the stimuli used to compute the RDMs. $^*P < 0.05$, $^{**}P < 0.001$, $^{****}P < 0.0001$ for difference from 0. Panels **a**–**c** adapted from ref. 33, US National Academy of Sciences; **d** and **e** adapted from ref. 35, S.M. Khaligh-Razavi and N. Kriegeskorte.

Though the top hidden layers of these goal-driven models end up being predictive of IT cortex data, they were not explicitly tuned to do so; indeed, they were not exposed to neural data at all during the training procedure. Models thus succeeded in generalizing in two ways. First, the models were trained for category recognition using real-world photographs of objects in one set of semantic categories, but were tested against neurons on a completely distinct set of synthetically created images containing objects whose semantic categories were entirely non-overlapping with that used in training. Second, the objective function being used to train the network was not to fit neural data, but instead the downstream behavioral goal (for example, categorization). Model parameters were independently selected to optimize categorization performance, and were compared with neural data only after all intermediate parameters—for example, nonlinear model layers—had already been fixed.

Stated another way, within the class of HCNNs, there appear to be comparatively few qualitatively distinct, efficiently learnable solutions to high-variation object categorization tasks, and perhaps the brain is forced over evolutionary and developmental timescales to pick such a solution. To test this hypothesis it would be useful to identify non-HCNN

Box 3 The meaning of 'understanding' in a complex sensory system

What does it mean to understand a complex neural system[67]? In this Perspective, we have suggested that successful models are image-computable, mappable and quantitatively predictive. But do models that meet these criteria necessarily represent understanding? It can be argued that deep neural networks are black boxes that give limited conceptual insight into the neural systems they aim to explain. Indeed, the very fact that deep HCNNs are able to predict the internal responses of a highly complex system performing a very nonlinear task suggests that, unlike earlier toy models, these deeper models will be more difficult to analyze than earlier models. There may be a natural tradeoff between model correctness and understandability.

Optimal stimulus and perturbation analysis. However, one of the key advantages of an image-computable model is that it can be analyzed in detail at low cost, making high-throughput 'virtual electrophysiology' possible. Recent techniques that optimize inputs either to match the statistics of target images or to maximize activation of a single output unit have produced impressive results in texture generation, image style matching and optimal stimulus synthesis (ref. 68 and Mordvintsev, A., Tyka, M. & Olah, C., http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html, 2015). These techniques could be used to identify the featural drivers of individual neurons, using the models' efficiency of scale to reduce a huge stimulus space to a set small enough to measure using realistic experimental procedures[69]. Inspired by causal intervention experiments[70], predictions for causal relationships between neural responses and behavior could be obtained by perturbing units within the model, even optimizing stimuli and perturbation patterns to achieve the most effective behavioral changes.

A concrete example of traversing Marr's levels of analysis. Goal-driven models yield higher level insight as well. That functional constraints can produce neurally predictive models is reminiscent of earlier work, including efficient coding hypotheses[23,24]. In both approaches, a driving concept—expressed as an objective function for optimization—explains why parameters are as they are. Unlike efficient coding, goal-driven HCNNs derive their objective function from behaviors that organisms are known to perform, rather than more abstract concepts, such as sparsity, whose ecological relevance is unclear. In this sense, the current work is more similar in spirit to Marr's levels of analysis[71], investigating how a system's computational-level goals influence its algorithmic and implementation level mechanisms. This approach is also related to neuroethology, where the natural behavior of an organism is studied to gain insight into underlying neural mechanisms[72].

models that, when optimized for categorization, achieved high performance. The hypothesis predicts that any such models would fail to predict neural response data.

Intermediate and lower layers predict V4 and V1 responses

In addition to higher model layers mapping to IT, intermediate layers of these same HCNN models turn out to be state-of-the-art predictors of neural responses in V4 cortex, an intermediate visual area that is the main cortical input to IT[33] (**Fig. 2c**). While the fit to IT cortex peaks in the highest hidden model layers, the fit to V4 peaks in the middle layers. In fact, these 'accidental' V4-like layers are significantly more predictive of V4 responses than models built from classical intuitions of what the area might be doing (for example, edge conjunction or curvature representation[37]). Continuing this trend, the lowest layers of goal-driven HCNN models naturally contain a Gabor-wavelet-like activation pattern. Moreover, these lower layers provide effective models of voxel responses in V1–V3 voxel data (**Fig. 2e**)[35,36]. Top-down constraints are thus able to reach all the way down the ventral hierarchy.

A common assumption in visual neuroscience is that understanding tuning curves in lower cortical areas (for example, edge conjunctions in V2 (ref. 38) or curvature in V4 (ref. 39)) is a necessary precursor to explaining higher visual areas. Results with goal-driven deep HCNNs show that top-down constraints can yield quantitatively accurate models of intermediate areas even when descriptive bottom-up primitives have not been identified (see **Box 3**).

HCNN layers as generative models of cortical areas. Unlike previous modeling approaches that fit single nonlinear models for each empirically measured neuron and then describe the distributions of parameters that were found[6], the performance-based approach generates a single model for all neurons simultaneously. Consequently, layers of the deep HCNNs are generative models for corresponding cortical areas, from which large numbers of (for example) IT-, V4- or V1-like units can be sampled. Given that the neurons used to evaluate model correctness were chosen by random electrode sampling, it is likely that any future neurons sampled from the same

Box 4 Gradient backpropagation

The basic idea of the gradient backpropagation algorithm is simple:
1. Formulate the task of interest as a loss function to be minimized—for example, categorization error. The loss function should be piecewise differentiable with respect to both the inputs (for example, images) and the model parameters.
2. Initialize the model parameters either at random or through some well-informed initial guess[14].
3. For each input training sample, compute the derivative of the error function with respect to the filter parameters, and sum these values over the input data.
4. Update network parameters by gradient descent—that is, by moving each parameter a small amount in the direction opposite to the error gradient for that parameter.
5. Repeat steps 3 and 4 until either the training error converges or, if overfitting is a concern, some 'early stopping' criterion is met[14].

The key insight that makes this procedure relatively efficient for feedforward networks is that—simply by applying the chain rule from basic calculus—the derivatives of the error with respect to filter values in a given layer can be efficiently computed from those in the layer just above[42]. Derivative computations thus start at the top layer and then propagate backwards through the network down to the first layers.

Another important technical innovation enabling large-scale backpropagation was stochastic gradient descent (SGD)[42]. SGD involves breaking training data into small, randomly chosen batches. Gradient descent is done on each batch in sequence until the training data are exhausted, at which point the procedure can begin again, usually on newly chosen random batches. SGD enables backpropagation on much larger data sets than previously contemplated and usually converges to a stable solution, though the statistical theory guaranteeing such convergence is not well developed.

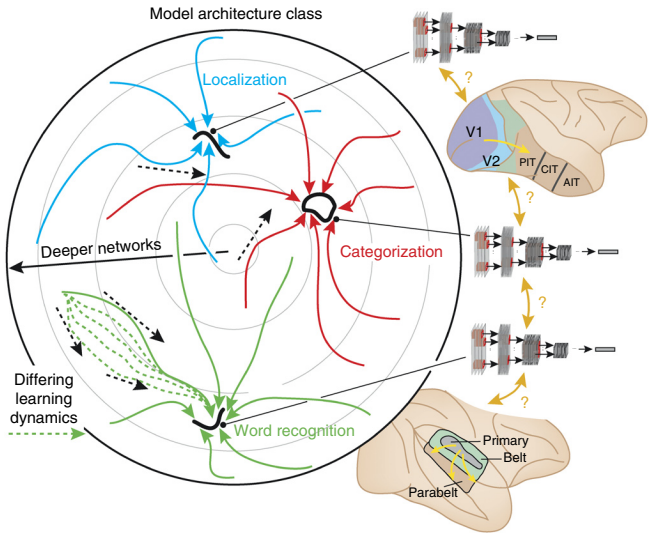

Figure 3 The components of goal-driven modeling. The large circle represents an architectural model class; each point in the space is a full model (examples at right); inner circles represent subspaces of the full model class containing models of a given number of layers. Goal-driven models are built by using learning algorithms (dotted black arrows) that drive systems along trajectories in the model class (solid colored lines) to discover especially optimal models. Each goal can be thought of as corresponding to a basin of attraction within the model class (thick black contours) containing parameters that are especially good for solving that goal. Computational results have shown that tasks put a strong constraint on model parameter settings, meaning that the set of optimal parameters for any given task is very small compared to the original space. These goal-driven models can then be evaluated for how predictive they are of the response properties of neurons in brain areas that are thought to underlie behavior in a given task domain. For example, the units of a model optimized for word recognition could be compared to response properties in the primary, belt and parabelt regions of auditory cortex[40]. Models can also be compared to each other to determine to what extent different types of tasks lead to shared neural structures. Various component rules (supervised, unsupervised or semi-supervised) can also be studied to determine how they might lead to different dynamics during postnatal development or expertise learning (dashed green paths).

areas will be equally well predicted, without having to update model parameters or train any new nonlinear functions.

Application to auditory cortex. A natural idea is to apply goal-based HCNN modeling to sensory domains that are less well understood than vision. The most obvious candidate for this is audition, where a clear path forward involves producing HCNN models whose top layers are optimized to solve auditory tasks such as speech recognition, speaker identification, natural sound identification and so on. An intriguing possibility is that intermediate layers of such models may reveal previously unknown structures in non-primary auditory cortex. Initial results suggest that this approach holds promise[40].

Factors leading to the improvement of HCNNs

Taking initial inspiration from neuroscience, HCNNs have become a core tool in machine learning. HCNNs have been successful on many tasks, including image categorization, face identification, localization, action recognition, depth estimation and a variety of other visual tasks[41]. Related recurrent versions of deep neural networks have been used to make strides in speech recognition. Here we discuss some of the technical advances that have led to this recent progress.

Hardware-accelerated stochastic error backpropagation for optimizing filter parameters

In supervised learning of a task (for example, car detection in images), one chooses a set of training data, containing both sample inputs (for example. images of cars and non-cars) and labels describing desired results for each input (for example, image category labels, such as "car" or "dog"). Learning algorithms are then used to optimize the parameter settings of the network so that output layers yield the desired labels on the training data[14]. A powerful algorithm for supervised learning of filter parameters from supervised data has been in existence for several decades: error gradient descent by backpropagation[14,42] (see **Box 4**). However, until recently, backpropagation has been computationally impractical at large scales on massive data sets. The recent advent of graphical processing unit (GPU)-accelerated programming has been a great boon because backpropagation computations largely involve either simple pointwise operations or parallel matrix dot-products[15,33,43]. GPUs, which are more neuromorphic than von Neumann CPU architectures, are especially well suited to these operations, routinely yielding speed increases of tenfold or more[15]. Further advances in neuromorphic computing could accelerate this trend[44].

Automated learning procedures for architectural parameters

Discrete architectural parameters (for example, number of layers) cannot easily be optimized by error backpropagation. However, discrete parameters are critical to final network performance[15,18]. Traditionally, these parameters had been chosen by hand, empirically testing various combinations one at a time until improvements were observed. More recently, procedures such as Gaussian process optimization and genetic algorithms have been deployed to learn better architectural parameters automatically[15,45,46].

Large web-enabled labeled data sets

Another important factor in recent advances is the advent of large labeled data sets. In the visual domain, early data sets often consisted of hundreds of images in hundreds of categories[47]. It was eventually realized that such data sets were neither large nor varied enough to provide sufficient training data to constrain the computational architecture[15,48]. A major advance was the release of the ImageNet data set, which contains tens of millions of images in thousands of categories, curated from the Internet by crowd-sourcing[49]. Taking advantage of these large data sets required the efficient hardware-accelerated algorithms described above. Once these were in place, much deeper neural networks could be trained. A rough rule of thumb is that the number of training samples for backpropagation should be 10 times the number of network parameters. Given that the number of parameters in a modern deep network far exceeds 100,000, the need for millions of training samples becomes evident, at least for current parameter learning strategies. (The neural learning algorithms used by the brain are probably significantly more efficient with labeled data than current computational methods for training HCNNs, and may not be subject to the '10×' heuristic.)

A concomitance of small tweaks to architecture class and training methods

A number of other small changes in neural network architecture and training helped improve performance. One especially relevant modification replaced continuously differentiable sigmoid activation functions with half-rectified thresholds[43]. Because these activation functions have constant or zero derivative almost everywhere, they

Box 5 Understanding adversarial optimization effects

An intriguing recent development in the exploration of HCNNs is the discovery of adversarial images: normal photographs that are subtly modified in ways that are undetectable to humans but that cause networks to incorrectly detect arbitrary objects in the modified image[73,74]. In effect, adversarial images demonstrate that existing HCNNs may be susceptible to qualitatively different types of illusions than those that fool humans. These images are created through adversarial optimization, a process in which the pixels of the original image are optimally modified so as to produce the largest changes in the network's final category-detection layer, but with the least disturbance at the pixel level. Creating such images, which may not naturally arise in the physical world, requires complete access to the network's internal parameters.

Thinking along the lines of three components of goal-driven modeling discussed above (and see **Fig. 3**), several possibilities for explaining adversarial examples include (i) that similar effects would be replicable in humans—for example, the creation of idiosyncratic images that fool one human but are correctly perceived by others—if experiments had access to the detailed microcircuitry of that individual brain and could run an adversarial optimization algorithm on it; (ii) that optimization for a categorization goal is brittle, but if richer and more robust optimization goal(s) were used, the effects would disappear; or (iii) that adversarial examples expose a fundamental architectural flaw in HCNNs as brain models, and only by incorporating other network structures (for example, recurrence) will the adversarial examples be overcome. Regardless of which (if any) if these is most correct, understanding adversarial optimization effects would seem to be a critical component of better understanding HCNNs themselves, especially as putative models of the brain.

suffer less from the so-called vanishing-gradients problem, in which error gradients in early layers become too small to optimize effectively. A second type of improvement was the introduction of regularization methods that inject noise during backpropagation into the network to prevent the learning of fragile, overfit weight patterns[43].

The unreasonable effectiveness of engineering

Recent improvements represent the accretion of a number of critical engineering improvements (for example, refs. 50,51). These changes may not signal major conceptual breakthroughs beyond the original HCNN and backpropagation concepts described decades ago, but they nonetheless led to enormous improvement in final results. Large data sets and careful engineering have been much more important than was originally anticipated[52].

Going forward: potentials and limitations

Goal-driven deep neural network models are built from three basic components (**Fig. 3**):

- a model architecture class from which the system is built, formalizing knowledge about the brain's anatomical and functional connectivity;
- a behavioral goal that the system must accomplish, such as object categorization; and
- a learning rule that optimizes parameters within the model class to achieve the behavioral goal.

The results above demonstrate how these three components can be assembled to make detailed computational models that yield testable predictions about neural data, significantly surpassing prior sensory cortical models. Future progress will mean, in part, better understanding each of these three components—as well as their limitations (see **Box 5**).

Improving architecture class

Continued success in using computational models to understand sensory cortex will involve more detailed and explicit mapping between model layers and cortical areas. HCNN operations such as template matching and pooling are neurally plausible, but understanding whether and how the parameterizations used in HCNNs actually connect to real cortical microcircuits is far from obvious. Similarly, while the hierarchy of HCNN model layers appears to generally correspond with the overall order of observed ventral cortical areas, whether the model-layer/brain-area match is one-to-one (or close to it) is far from fully understood. Recent high-performing computer vision networks have greatly increased the number of layers, sometimes to 20 or more[50]. Evaluating whether these very deep networks are better explanations of neural data will be of importance, as deviations from neural fit would suggest that the architectural choices are different from those in the brain. More generally, one can ask, within the class of HCNNs, which architectures, when optimized for categorization performance, best fit the ventral steam neural response data? The results above argue that this could be a new way to infer the architectures in the adult ventral stream.

Such top-down, performance-driven approaches should of course be coupled with state-of-the-art experimental techniques such as two-photon microscopy, optogenetics, electron microscopy reconstruction and other tracing techniques that aim to narrow the class of architectures more directly. Better empirical understanding at the neural circuit level could allow a narrowing in the class of biologically relevant HCNNs, ruling out certain architectures or making informed initial guesses about filter parameters. Models would then need to learn fewer parameters to achieve equal or better neural predictivity.

In both vision and audition, model architecture class could also be improved by building more biologically realistic sensor front-ends into early layers, using known results about subcortical structures[53]. At the opposite end of the scale spectrum, there are large-scale spatial inhomogeneities in higher cortical areas (for example, face patches)[4]. In the lower layers of HCNNs, there is an obvious mapping onto the cortical surface via retinotopic maps, but this relationship is less clear in higher layers. Understanding how multidimensional deep network output may map to two-dimensional cortical sheets, and the implications of this for functional organization, are important open problems.

Improving goal and training-set understanding

The choice of goal and training set has significantly influenced model development, with high-variation data sets exposing the true heterogeneity within real-world categories[33,48,49]. It seems likely that this data-driven trend will continue[52]. A key recent result is that HCNNs trained for one task (for example, ImageNet classification) generalize to many other visual tasks quite different from the one on which they were originally trained[41]. If many relevant tasks come along 'for free' with categorization, which tasks do not? An especially important open challenge is finding tasks that are not solved by categorization optimization but rather require direct independent optimization, and then testing models optimized for these tasks to see if they better explain ventral stream neural data. Developing rich new labeled data sets will be critical to this goal. Understanding how HCNNs systems for various sensory tasks relate to each other, in terms of shared or

divergent architectures, would be of interest, both within a sensory domain[54], as well as across domains (for example, between vision and audition; see **Fig. 3**).

Improving learning rule understanding

While it is valuable that supervised learning creates working models that are a remarkably good fit to real perceptual systems, it is physiologically unlikely that cortex is implementing exact backpropagation. A core inconsistency between current deep-learning approaches and real biological learning is that training effective HCNNs requires very large numbers of high-level semantic labels. True biological postnatal learning in humans, higher primates and other animals may use large amounts of unsupervised data, but is unlikely to require such large amounts of externally labeled supervision. Discovering a biologically realistic unsupervised or semi-supervised learning algorithm[55–57] that could produce high levels of performance and neural predictivity would be of interest, from both artificial intelligence and neuroscience viewpoints.

Beyond sensory systems and feedforward networks

Largely feedforward HCNNs cannot provide a full account of dynamics in brain systems that store extensible state, including any that involve working memory, since the dynamics of a feedforward network will converge to the same state independent of input history. However, there is a growing body of literature connecting recurrent neural networks to neural phenomena in attention, decision making and motor program generation[58]. 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 for cases in which there is a complex loop between behavioral output and input stimulus—for example, when modeling exploration of an agent over long time scales in a complex sensory environment[59]. Intriguing recent results from reinforcement learning[60] have shown how powerful in solving strategy-learning problems deep neural network techniques may be. Mapping these to ideas in the neuroscience of the interface between ventral visual cortex and, for example, parietal cortex or the hippocampus will be of great interest[61,62].

Conclusion

In sum, deep hierarchical neural networks are beginning to transform neuroscientists' ability to produce quantitatively accurate computational models of the sensory systems, especially in higher cortical areas where neural response properties had previously been enigmatic. Such models have already achieved several notable results, explaining multiple lines of neuroscience data in both humans and monkeys[33–36]. However, like any scientific advance of importance, these ideas open up as many new questions as they answer. There is much exciting and challenging work to be done, requiring the continued rich interaction between neuroscience, computer science and cognitive science.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. DiCarlo, J.J. & Cox, D.D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
2. DiCarlo, J.J., Zoccolan, D. & Rust, N.C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
3. Felleman, D.J. & Van Essen, D.C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
4. Malach, R., Levy, I. & Hasson, U. The topography of high-order human object areas. *Trends Cogn. Sci.* **6**, 176–184 (2002).
5. Carandini, M. *et al.* Do we know what the early visual system does? *J. Neurosci.* **25**, 10577–10597 (2005).
6. Sharpee, T.O., Kouh, M. & Reynolds, J.H. Trade-off between curvature tuning and position invariance in visual area V4. *Proc. Natl. Acad. Sci. USA* **110**, 11618–11623 (2013).
7. David, S.V., Hayden, B.Y. & Gallant, J.L. Spectral receptive field properties explain shape selectivity in area V4. *J. Neurophysiol.* **96**, 3492–3505 (2006).
8. Gallant, J.L., Connor, C.E., Rakshit, S., Lewis, J.W. & Van Essen, D.C. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J. Neurophysiol.* **76**, 2718–2739 (1996).
9. Rust, N.C., Mante, V., Simoncelli, E.P. & Movshon, J.A. How MT cells analyze the motion of visual patterns. *Nat. Neurosci.* **9**, 1421–1431 (2006).
10. Hubel, D.H. & Wiesel, T.N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol. (Lond.)* **148**, 574–591 (1959).
11. Fukushima, K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
12. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
13. Serre, T., Oliva, A. & Poggio, T. A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. USA* **104**, 6424–6429 (2007).
14. Bengio, Y. *Learning Deep Architectures for AI* (Now Publishers, 2009).
15. Pinto, N., Doukhan, D., DiCarlo, J.J. & Cox, D.D. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput. Biol.* **5**, e1000579 (2009).
16. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. in *The Handbook of Brain Theory and Neural Networks* 255–258 (MIT Press, 1995).
17. Carandini, M. & Heeger, D.J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
18. Yamins, D., Hong, H., Cadieu, C. & Dicarlo, J. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. *Adv. Neural Inf. Process. Syst.* **26**, 3093–3101 (2013).
19. De Valois, K.K., De Valois, R.L. & Yund, E.W. Responses of striate cortex cells to grating and checkerboard patterns. *J. Physiol. (Lond.)* **291**, 483–505 (1979).
20. Jones, J.P. & Palmer, L.A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1233–1258 (1987).
21. Movshon, J.A., Thompson, I.D. & Tolhurst, D.J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol. (Lond.)* **283**, 53–77 (1978).
22. Klein, D.J., Simon, J.Z., Depireux, D.A. & Shamma, S.A. Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex. *J. Comput. Neurosci.* **20**, 111–136 (2006).
23. Barlow, H.B. Possible principles underlying the transformations of sensory messages. in *Sensory Communication* Vol. 1, 217–234 (1961).
24. Olshausen, B.A. & Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
25. deCharms, R.C. & Zador, A. Neural representation and the cortical code. *Annu. Rev. Neurosci.* **23**, 613–647 (2000).
26. Olshausen, B.A., Sallee, P. & Lewicki, M.S. Learning sparse image codes using a wavelet pyramid architecture. *Adv. Neural Inf. Process. Syst.* **14**, 887–893 (2001).
27. Logothetis, N.K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).
28. Zoccolan, D., Kouh, M., Poggio, T. & DiCarlo, J.J. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* **27**, 12292–12307 (2007).
29. Kriegeskorte, N. Relating population-code representations between man, monkey, and computational models. *Front. Neurosci.* **3**, 363–373 (2009).
30. Ullman, S. Visual routines. *Cognition* **18**, 97–159 (1984).
31. Singer, W. & Gray, C.M. Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* **18**, 555–586 (1995).
32. 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. *J. Neurosci.* **35**, 13402–13418 (2015).
33. Yamins, D.L.K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111**, 8619–8624 (2014).
34. Cadieu, C.F. *et al.* Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10**, e1003963 (2014).
35. Khaligh-Razavi, S.M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
36. Güçlü, U. & van Gerven, M.A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
37. Yau, J.M., Pasupathy, A., Brincat, S.L. & Connor, C.E. Curvature processing dynamics in macaque area V4. *Cereb. Cortex* **23**, 198–209 (2013).
38. Freeman, J. & Simoncelli, E.P. Metamers of the ventral stream. *Nat. Neurosci.* **14**, 1195–1201 (2011).

39. Pasupathy, A. & Connor, C.E. Population coding of shape in area V4. *Nat. Neurosci.* **5**, 1332–1338 (2002).
40. Kell, A., Yamins, D., Norman-Haignere, S. & McDermott, J. Functional organization of auditory cortex revealed by neural networks optimized for auditory tasks. *Soc. Neurosci. Abstr.* 466.04 (2015).
41. Razavian, A.S., Azizpour, H., Sullivan, J. & Carlsson, S. CNN features off-the-shelf: an astounding baseline for recognition. in *Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Conference on*, 512–519 (IEEE, 2014).
42. Bottou, L. Large-scale machine learning with stochastic gradient descent. in *Proc. COMPSTAT 2010*, 177–186 (Springer, 2010).
43. Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
44. Choudhary, S. *et al.* Silicon neurons that compute. in *Artificial Neural Networks and Machine Learning–ICANN 2012*, 121–128 (Springer, 2012).
45. Snoek, J., Larochelle, H. & Adams, R.P. Practical bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **26**, 2951–2959 (2012).
46. Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proc. 30th International Conference on Machine Learning* 115–123, http://jmlr.csail.mit.edu/proceedings/papers/v28/ (2013).
47. Griffin, G., Holub, A. & Perona, P. The Caltech-256 object category dataset. Caltech Technical Report, http://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001 (2007).
48. Pinto, N., Cox, D.D. & DiCarlo, J.J. Why is real-world visual object recognition hard? *PLoS Comput. Biol.* **4**, e27 (2008).
49. Deng, J. *et al.* ImageNet: a large-scale hierarchical image database. in *CVPR 2009, IEEE Conference on Computer Vision and Pattern Recognition*, 248–288 (IEEE, 2009).
50. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at http://arxiv.org/abs/1409.1556 (2014).
51. Szegedy, C. *et al.* Going deeper with convolutions. Preprint at http://arxiv.org/abs/1409.4842 (2014).
52. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
53. Pillow, J.W. *et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
54. Khorrami, P., Paine, T.L. & Huang, T.S. Do deep neural networks learn facial action units when doing expression recognition? Preprint at http://arxiv.org/abs/1510.02969 (2015).
55. Hinton, G.E., Dayan, P., Frey, B.J. & Neal, R.M. The "wake-sleep" algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).
56. Zhu, L.L., Lin, C., Huang, H., Chen, Y. & Yuille, A. Unsupervised structure learning: hierarchical recursive composition, suspicious coincidence and competitive exclusion. in *Computer Vision–ECCV 2008*, 759–773 (Springer, 2008).
57. Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In *Unsupervised and Transfer Learning: Challenges in Machine Learning* Vol. 7 (eds. Guyon, I., Dror, G & Lemaire, V.) 29–41 (Microtome, 2013).
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. Stadie, B.C., Levine, S. & Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. Preprint at http://arxiv.org/abs/1507.00814 (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).
62. Hulbert, J. & Norman, K. Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cereb. Cortex* **25**, 3994–4008 (2015).
63. 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).
64. 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**, 12978–12995 (2010).
65. 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).
66. Pagan, M., Urban, L.S., Wohl, M.P. & Rust, N.C. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nat. Neurosci.* **16**, 1132–1139 (2013).
67. Marder, E. Understanding brains: details, intuition, and big data. *PLoS Biol.* **13**, e1002147 (2015).
68. Gatys, L.A., Ecker, A.S. & Bethge, M. A neural algorithm of artistic style Preprint at http://arxiv.org/abs/1508.06576 (2015).
69. 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.* **11**, 1352–1360 (2008).
70. Afraz, A., Boyden, E.S. & DiCarlo, J.J. Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proc. Natl. Acad. Sci. USA* **112**, 6730–6735 (2015).
71. Marr, D., Poggio, T. & Ullman, S. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information* (MIT Press, 2010).
72. Hoyle, G. The scope of neuroethology. *Behav. Brain Sci.* **7**, 367–381 (1984).
73. Szegedy, C. *et al.* Intriguing properties of neural networks. Preprint at http://arxiv.org/abs/1312.6199 (2013).
74. Goodfellow, I.J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. Preprint at http://arxiv.org/abs/1412.6572 (2014).

- title: Using goal-driven deep learning models to understand sensory cortex
- year: 2016
- author: Daniel L K Yamins and James J DiCarlo
- journal: Nature Neuroscience
- volumne: 19
- Number: 3
- pages: 356-365

ゴール駆動型深層学習モデルを用いた感覚皮質の理解

要旨

コンピュータビジョンや人工知能の分野における技術革新を背景に， 計算論的神経科学の分野では， ゴール駆動型の階層型畳み込みニューラルネットワーク (HCNN) を用いて，高次視覚皮質領域における神経の単一ユニットおよび集団反応のモデル化が進展している。 この展望では， 最近の進歩をより広範なモデリングの文脈でレビューし， それを支えた主要な技術革新について説明する。 そして， ゴール駆動型 HCNN のアプローチが， 感覚皮質処理の発達と組織化をより深く理解するためにどのように利用できるかを説明する。

1. 感覚皮質のモデルに何を期待すべきか？

脳は， 入力された感覚データを， 宿主である生物の行動上の必要性に応じて， 積極的に再構成する (図1a)。 人間の視覚では， 網膜からの入力が物体中心の豊かな情景に変換される。 人間の聴覚では， 音波が言葉や文章に変換される。 問題の核心は， 感覚入力空間の自然な軸 (例えば， 光受容体や有毛細胞の電位) が， 行動に関連する高レベルの構成要素が変化する軸とうまく整合していないことである。 例えば， 視覚データでは， 物体の移動， 回転， 奥行き方向の動き， 変形， 照明の変化などにより， 元の入力空間 (網膜) に複雑な非線形変化が生じる。 逆に， 生態学的には全く異なる 2 つの物体 (例えば， 異なる個人の顔) の画像が， 画素空間では非常に近い位置にあることもある。 このように，行動に関連する次元は入力空間に「絡み合って」おり， 脳はその絡み合いを解きほぐさなければならない(1, 2)。

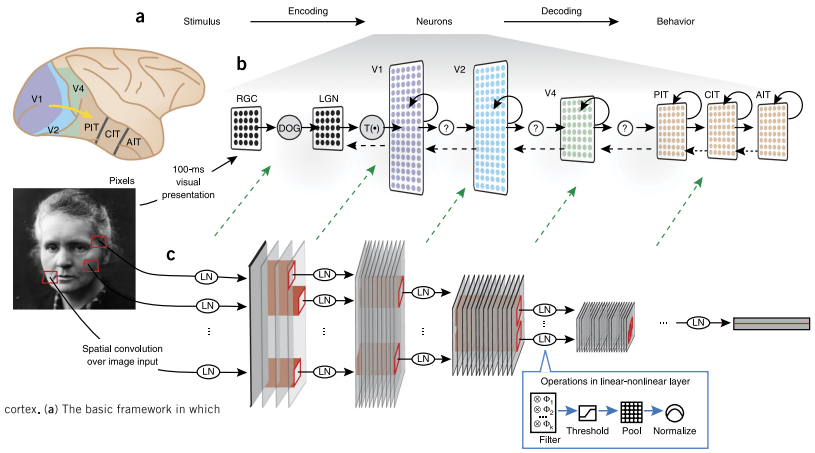

図1 感覚皮質のモデルとしてのHCNN。 (a) 感覚野を研究する基本的な枠組みは， 刺激が神経活動のパターンに変換される過程である「符号化」と， 神経活動が行動を生み出す過程である「復号化」である。 HCNN は， 刺激が脳内で測定された神経反応にマッピングされるという， 符号化ステップのモデルを作るのに使われている。 (b) 腹側視覚経路は， 最も包括的に研究されている感覚カスケードである。 感覚カスケードは， 一連の皮質脳領域の連結で構成されている (マカクサルの脳を示す)。 PIT は後下側頭皮質。 CIT 中央。 AIT 前部。 RGC 網膜神経節細胞。 LGN 外側膝状核。 DoG, ガウシアン差分モデル。 T() 変換。 (c) HCNN は多層構造のニューラルネットワークで， 各層はフィルタリング， 閾値， プーリング， 正規化などの単純な演算を LN (Linear-Nonlinear) で組み合わせて構成されている。 各層のフィルターバンクは， シナプスの強さに類似した重みのセットで構成されている。 フィルターバンクの各フィルターは， 異なる周波数と方向性を持つガボールウェーブレットに類似した， 明確なテンプレートに対応している。 層内の演算は， 入力内の空間的なパッチに局所的に適用され， 単純で限られたサイズの受容野 (赤枠) に対応する。 複数の層を構成することで， 元の入力刺激が複雑な非線形変換を受けることになる。 各層では， レチノピーが減少し， 有効な受容野サイズが増加する。 HCNN は， 腹側視覚経路のモデルに適した候補である。 HCNN は， 定義上， 画像計算可能であり， 任意の入力画像に対する応答を生成することができる。 HCNN はまた， マッピング可能である。 つまり， 腹側視覚経路の観測可能な構造と，コンポーネントごとに自然に識別できる。 また， パラメータが正しく選択されていれば， 予測可能である。 つまり、ネットワーク内の各層は， モデルが構築された領域の外にある大規模なクラスの刺激に対する神経応答パターンを記述する。

大脳皮質の感覚システムは， 解剖学的に区別されながらもつながっている一連の領域 (3,4)で構成されていること(図1b) と， 刺激変化後の最初の 100 msにおける神経活動の初期波は， その一連の領域に沿ってカスケードのように展開すること (2) という， 2 つの基本的な経験則がある。 カスケードの個々のステージでは， 入力の加重線形和や， 活性化閾値や競合正規化などの非線形性など， 非常に単純な神経演算が行われる(5)。 しかし， 単純なステージを直列に適用すると， 複雑な非線形変換が生じることがある(6)。 もともとの入力のもつれが非常に非線形であるため， もつれを解くプロセスも非常に非線形でなければならない。

脳のニューラルネットワークが計算できる非線形変換の可能性は膨大である。 感覚システムを理解する上での大きな課題は， システムの識別である。 つまり， 真の生体回路がどのような変換を行っているのかを特定することである。 神経伝達関数の概要 (受容野の特徴など) を把握することは有用だが (7)， このシステム同定の問題を解決するには， 最終的には符号化モデルを作成する必要がある。 すなわち， 任意の刺激 (例えば， 任意の画素地図) を入力し， その刺激に対する神経反応を正しく予測して出力するアルゴリズムである。 モデルは， 厳選されたニューロンで確認された狭い現象を， 高度に制御された単純な刺激に対してのみ定義して説明するだけのものではない(8,9)。 任意の刺激で動作することと， ある領域内のすべてのニューロンの反応を定量的に予測することは， 感覚領域のモデルが満たすべき 2 つの中核的な基準である (ボックス 1 参照)。

Box 1 感覚的なエンコーディングモデルの最低限の基準

我々は， 感覚皮質システムの符号化モデルが満たすべき 3 つの基準を特定する。 刺激の計算可能：モデルは対象とする一般的な刺激領域内の任意の刺激を受け入れるべきである。 対応付け可能性：モデルの構成要素が， 実験的に定義可能な神経系の構成要素に対応していること。 予測可能性： モデルのユニットは， マッピングされた各領域内の任意に選択されたニューロンについて， 刺激ごとの反応を詳細に予測する必要がある。 これらの基準は時に相反することがある。 最も細かい粒度でのマッピング可能性にこだわると， 複雑な実世界の刺激に対して実際に機能するモデルの特定を妨げる可能性がある。 簡略化された文脈で神経回路の接続性の詳細なモデルを求めることは重要だが， そのようなモデルが全体として， 実世界の刺激に対する神経反応の正確な予測につながらなければ， 低レベルの妥当性の有用性は限られてしまう。

2. 階層型畳み込みニューラルネットワーク

Hubel と Wiesel (10) の画期的な研究に始まり， 視覚システム神経科学の研究は， 脳が階層的に組織された一連の皮質領域， 腹側視覚路を介して不変の物体認識行動を生成することを示してき (2)。 Hubel と Wiesel のアイデアを一般化して， 生物学的にインスパイアされたニューラルネットワークを構築した研究者も数多くいる (例えば， 文献 11-15)。 やがて， これらのモデルは， HCNN と呼ばれる， より一般的なクラスの計算アーキテクチャの例であることがわかってきた(16)。 HCNNは， 感覚入力に繰り返される単純な神経回路モチーフを含む層を積み重ね， それらの層を直列に構成する。 各層は単純だが， このような層で構成された深いネットワークは， 入力データの複雑な変換を計算する。 これは腹側経路で生成される変換に類似している。

3. HCNN の各層に含まれるモチーフ

HCNN の 1 つの層を構成する具体的な演算は， 広く観察されている LN (linear-onlinear) 神経モチーフ (5) にヒントを得ている。 これらの演算 (図1c) には，

- (i) 入力刺激の局所的なパッチと一連のテンプレートとのドット積をとる線形演算であるフィルタリング，
- (ii) 点ごとの非線形演算である活性化 (典型的には， 整流された線形閾値またはシグモイド)，
- (iii) 非線形の集約演算であるプーリング (典型的には， 局所的な値の平均または最大値) ，
- (iv) 単一の HCN 層を構成する具体的な演算が含まれる(13)
- (iv) 分割正規化: 出力値を標準的な範囲に補正する(17) すべての HCNN がこれらの演算をこの順番で使用しているわけではないが， ほとんどが似通っている。 すべての基本的な演算は， HCNN の 1 つの層に存在し， その層は， 通常， 1 つの皮質領域にマッピングされる。

神経の受容野と同様， HCNN のすべての演算は， 入力の空間的広がりよりも小さい固定サイズの入力領域に， 局所的に適用される (図1c)。 例えば， 256×256 画素の画像の場合， 1 つの層の受容野は 7×7 画素になります。 このように空間的に重なり合っているため， フィルタリングやプーリングの操作は一般的に「ストライド」と呼ばれ， 各空間次元のほんの一部の位置でのみ出力が保持される。 画像の畳み込みでストライドが 2 の場合， 2 行目と 2 列目をスキップすることになる。

HCNN では， フィルタリングは畳み込み重み共有によって実装されており， すべての空間位置で同じフィルタテンプレートが適用される。 すべての場所で同じ演算が適用されるので， 出力の空間的変化は， 入力刺激の空間的変化に完全に起因する。 脳が文字通り重み付けを行っているとは思えない。 腹側視覚路や他の感覚皮質の生理学的性質から， 共有テンプレートが保存される単一のマスターロケーションの存在は否定されているようである。 しかし， 世界の自然な視覚 (または聴覚) 統計は， それ自体が空間 (または時間) においてほとんど変化しないので， 脳内の経験に基づく学習プロセスは， 異なる空間 (または時間) 位置の重みを収束させる傾向があるはずである。 したがって， 共有された重みは， 少なくとも中心視野内では， 脳の視覚システムを合理的に近似していると考えられる。 実際の視覚システムには強い焦点バイアスがあり， 不均一な受容野密度をより現実的に扱うことで， モデルの神経データへの適合性が向上するかもしれない。

4. スタッキングによるディープネットワーク

畳み込み層の出力は， 入力と同じ空間配置を持つため， ある層の出力を別の層に入力することができる。 そのため， HCNN を積み重ねることで， 深いネットワークを構築することができる (図1c)。 1 つの層のユニットが見る局所的な場の大きさは固定されていて小さいが， 元の入力に対する有効な受容野の大きさは， 層を重ねるごとに大きくなる。 これは経験的な観察結果と一致する (4)。 しかし， 各層で使用されるフィルターテンプレートの数は， 通常増加する。 しかし， 各層で使用されるフィルターテンプレートの数は一般的に増加し， 広く浅い層から深く狭い層へと次元が変化していく (図1c)。 多くの層を重ねると， 出力の空間成分が減少して畳み込みが意味をなさなくなるため， 1 つ以上の完全連結層を用いてネットワークを拡張するのが一般的である。 例えば， 複数の視覚カテゴリーのそれぞれについて， 入力画像にそのカテゴリーの物体が含まれている可能性を 1 つの出力ユニットで表現することができる。

5. パラメータ化されたモデル群としての HCNN

HCNN は単一のモデルではなく， パラメータ化されたモデル群である。 HCNN の特徴は以下の通りである。

- ネットワークに含まれる層の数を含む離散的なアーキテクチャ・パラメータと， 各層について，フィルタ・テンプレートの数， 各フィルタリング， プーリング， 正規化操作の局所的な半径， プーリングの種類を指定する離散的なパラメータ， かつ HCNN の実装に必要なその他の選択肢を指定する。
- 畳み込み層と完全連結層のフィルタの重みを指定する連続フィルタのパラメータ。

パラメータの選択は， 一見すると些細なことのように思えるが， 微妙なパラメータの違いが， 認識課題におけるネットワークの成績や， ニューラルデータとのマッチングに劇的な影響を与える (15,18)。

ボックス1 で述べた最小限のモデル基準があれば， 重要な目標は， 層が対象となる皮質システム内の異なる領域(例えば， 腹側経路の異なる領域) に対応し， それらの領域の反応パターンを正確に予測する単一の HCNN パラメータ設定を特定することである (ボックス2)。

> Box 2 モデルと神経感覚システムのマッピング
>
> 人工ニューラルネットワークを実際のニューロンにマッピングするにはどうすればよいのか。神経の詳細レベルに応じて，いくつかのアプローチが可能である。
>
> **タスク情報の一貫性** 最も粗いレベルでは， モデルがシステムに類似しているかどうかの有用な指標は， 潜在的な行動課題をサポートするために利用可能な明示的に復号可能な情報のパターンの一貫性である。 この方法では， モデルの「ニューロン」集団と， 記録されたニューロン集団を， 高レベルの課題 (例えば， 物体認識、顔識別など) のバッテリーについて， 同一の復号化方法で分析する。 必須ではないが， 線形分類器や線形回帰器 (1,32,63,64) などの単純な復号化を使用することは， 下流の復号化回路を仮想的に具現化する上で有用である (65, 66)。 このようにして， モデルと神経集団の両方に応答選択のパターンを生成する。 これらのパターンは， 粗い粒度 (例えば， 課題ごとの精度レベル (32)) または細かい粒度 (刺激ごとの反応の一貫性) で互いに比較される。 このアプローチは， モデルとニューロンの両方を， 動物や人間の被験者から得られた行動測定値と比較することができるため， ニューロン集団と行動の関連性に自然に結びつくことを指摘しておく(32)。 行動に最も直結していると考えられる神経領域 (例えば， 視覚の場合は IT) と， その領域の計算モデルの両方が， それらの行動パターンと高い整合性を示す必要がある(32)。
>
> **母集団の表現の類似性** もう 1 つの母集団レベルの指標として， 表現の類似性分析 (29,35) がある。 この分析では， 2 つの表現 (実際のニューロンの表現とモデルの表現) を， 対となる刺激の相関行列で特徴づける(図2d)。 この行列は， 与えられた刺激のセットに対して， 表現が「考える」刺激の各対がどれだけ離れているかを表している。 実際の神経集団の表現がそうであるように， モデルが刺激のペアを互いに近い (または遠い) ものとして扱う場合， モデルは神経表現に似ていると判断される。
>
> **シングルユニットの応答予測性** モデルのニューロンへのより詳細なマッピングとして， シングルユニットの線形神経応答予測性がある (33)。 この考え方は， 簡単な思考実験で理解することができる。 例えば， 2 匹の動物のある脳領域のすべてのニューロンの測定値があるとする。 ソース動物とターゲット動物です。 ソースのニューロンとターゲットのニューロンをどのようにマッピングするだろうか？ 多くの脳領域 (例えば， V4 や IT など) では， 動物間のユニットの正確な一対一のマッピングはないかもしれない。 しかし， 2 つの動物の領域は， 線形変換までは同じ (または非常に似ている) と考えるのが妥当である。 例えば， ターゲットの動物のユニットは， ソースの動物の (少数の) ユニットのほぼ線形結合であると考えられる。 工学的に言えば， 2 つの動物は感覚表現の「等価ベース」であると言える。 (もしマッピングが非線形でなければならないとしたら， そもそも 2 つの領域が動物間で同じであるかどうかが問題となる)。 マッピングを行うということは， 実質的には， 正しい線形の組み合わせを特定するという問題になる。 同様の考え方で， モデル層のユニットと

脳領域のニューロンを対応させることができる。 具体的には， 経験的に測定された各ニューロンを， モデル層のユニットからの線形回帰の対象として扱う。 目標は， モデルユニットの線形結合を見つけて， 元々の対象となる実在のニューロンと同じ応答パターンを確実に持つ「合成ニューロン」を作り出すことである:

find $c_i, i\in\left{1,\ldots,n\left}$ such that:

$$r(x) \simeq r_{synth}(x) = \sum_j c_j m_j(x)$$

ここで， $r(x)$ は刺激 x に対するニューロン r の応答， $m_i(x)$ は i 番目のモデルユニット (固定されたモデル層) の応答である。 r_{synth} の精度は, 係数 c_i の同定に使用されていない新しい刺激に対する r の説明された分散（R^2）として測定される。 理想的には， 非ゼロの重み c_i を持つモデルソースユニットiの数は， ある動物のニューロンを別の動物の同じ脳領域のニューロンにマッピングしようとしたときに経験的に見出される数とほぼ同じになる。

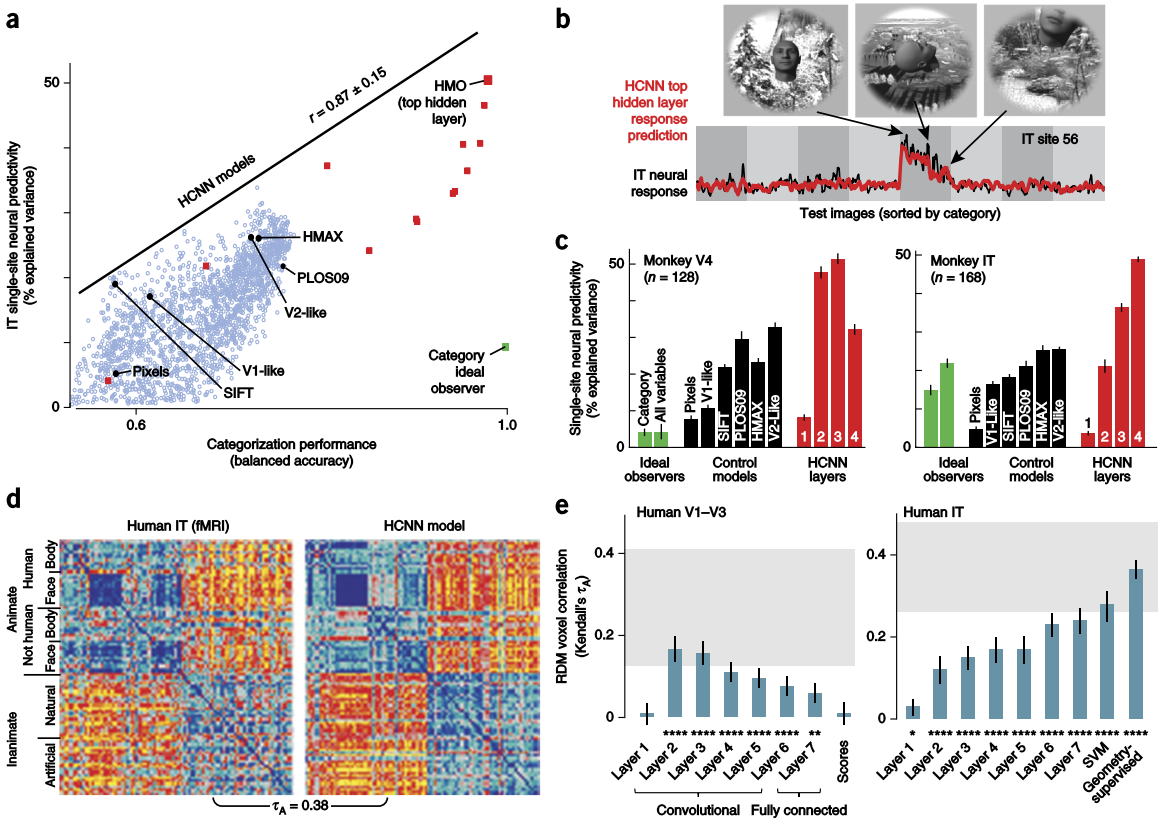

図2 目的に応じた最適化により， 腹側視覚野の神経学的予測モデルが得られる

(a)物体の分類を解くために最適化された HCNN モデルは， IT 神経応答の分散を予測するのに適した隠れ層表現を生み出す。 y 軸は，HCNN モデルの最後の隠れ層の IT 反応予測能力の中央値を $n = 168$ の IT 部位について示している。 部位応答は， 画像開始後 70〜170 ms後の平均発火率として定義されている。 応答予測性は Box 2 のように定義されている。 各ドットは， 大規模な HCNN モデル群の中から選ばれたモデルである。 青色の円で示されたモデルは， 物体分類の性能最適化からランダムに選ばれたものである。 黒丸は， コントロールモデルと， それ以前に発表された HCNN モデル。 赤色の四角は, 特定の HCNN モデルを生成する最適化手順で生成された HCNN モデルの経時変化を示す(33)。 PLOS09, ref. 15; SIFT, shape-invariant feature transform; HMO, optimized HCNN.

(b) 1 つの IT 神経部位に対する HCNN モデルの最後の隠れ層のモデル予測値 (赤のトレース) に対する実際の神経応答(黒のトレース)。 x 軸は $1,600$ 枚のテスト画像を示しているが， いずれもモデルの適合には使用されていない。 画像は， まずカテゴリーの同一性でソートされ， 次に変化量でソートされる。 各カテゴリーブロック内では， 右に向かってより急激な画像変換が行われている。 Y 軸は， 各テスト画像に対する神経部位の反応とモデル予測を表している。 この部位の反応には顔の選択性が見られましたが (挿入画像参照)， 予測性の結果は他の IT 部位でも同様でした(33)。

(c) 様々なモデルに対する IT と V4 の単一部位の神経反応予測性の比較。 予測率の中央値を示す棒の高さは V4 の 128 個の予測ユニット (左パネル) または IT の 168 個のユニット (右パネル) で取ったものです。 HCNN モデルの最後の隠れた層が IT の反応を最もよく予測し， 最後から 2 番目の隠れた層が V4 の反応を最もよく予測している。

(d) 人間の IT と HCNN モデルの代表的非類似度行列 (RDM)。 青色は低い値を示し， 表現は画像ペアを類似したものとして扱い， 赤色は高い値を示し， 表現は画像対を異なるものとして扱う。 値の範囲は 0 から 1 である。

(e) HCNN モデルの層の特徴と， 人間の V1-V3 (左)， 人間の IT (右) との間の Kendall's τA で 測定した RDM 類似度。 灰色の横棒は， ノイズや被験者間のばらつきを考慮した場合の真のモデルの性能の範囲を示す。 エラーバーは， RDM の計算に使用した刺激のブートストラップ・リサンプリングによって推定された s.e.m. である。 *P < 0.05, **P < 0.001, ****P < 0.0001 (0との差)。 パネル a-c は 文献より引用。文献 33, US National Academy of Sciences; d と e は文献 35, S.M. Khaligh-Razavi and N. Kriegeskorte.

過度の単純化だが， フィルタの変更とアーキテクチャのパラメータの関係は， 発達段階の変化と進化段階の変化の関係に類似している。 フィルタのパラメータはシナプスの重みに相当すると考えられ， その学習アルゴリズム (下記のバックプロパゲーションの説明を参照) はオンライン方式でパラメータを更新する。 一方， アーキテクチャのパラメータを変更すると， 計算プリミティブ， 感覚領域 (モデル層) の数， 各領域のニューロンの数が再構成される。

6. 視覚野の初期モデルの状況について

生体システムに最適な HCNN のパラメータを特定するために， さまざまなアプローチがとられてきた。

6.1 Hubel と Wiesel 理論によるパラメータの手作業でのデザイン

HCNN の概念が確立される以前の 1970 年代から， モデラーたちは， 比較的浅いネットワークでニューロンを説明できる可能性のある V1 などの大脳皮質下部領域に取り組み始めた。 Hubel と Wiesel の経験的な観察によると， V1 のニューロンはガボールウェーブレットフィルターに似ており， 異なるニューロンが異なる周波数と方向のエッジに対応していることが示唆された (10, 19)。 実際， 初期の計算モデルでは， 手で設計したガボールフィルターバンクを畳み込み重みとして使用し， V1 の神経反応を説明することに成功し

た (20)。 その後，閾値， 正規化， ゲインコントロール などの非線形性を利用することで， モデルを大幅に改善できることがわかり (5, 21)， HCNN クラスの最初の動機付けとなった。 同様の考え方は， 一次聴覚野のモデル化にも提案されている (22)。

6.2 効率的なコーディング制約によるパラメータの学習

Barlow, Olshausen らの研究により， フィルタのパラメータを決定する別の方法が導入された (23, 24)。 フィルタは， 元の入力を再構成する能力を維持しながら， 任意の刺激によって活性化されるユニットの数を最小化するように最適化された。 このような「スパース」で効率的なコーディングは， 自然の画像データからガボールウェーブレットのようなフィルターを自然に学習し， それらのパターンを手作業で構築する必要がない。

6.3 神経データへのネットワークの適合

1990 年代半ばに始まったもう一つの自然なアプローチは， 神経科学のデータをモデルのパラメータ選択に直接反映させることであった。 これは， 興味のある脳領域のニューロンの様々な刺激に対する応答データを収集し， 統計的フィッティング技術を用いて， 観察された刺激-応答関係を再現するモデルパラメータを見つけるというものである。 この戦略は， 視覚野 V1， 聴覚野 A1， 体性感覚野 S1 に浅いネットワークを当てはめることに成功した (文献25)。

6.4 より深いネットワークの難しさ

初期の皮質領域の浅い畳み込みモデルが成功したならば， より深いモデルが下流の感覚領域に光を当ててくれるかもしれない。 しかし， そのような高次の領域をモデル化するために必要な深いモデルは， V1 のようなモデルよりも多くのパラメータを持つことになる。 これらのパラメータはどのように選択すればよいのだろうか？

高次層で動作する出力は可視化が難しく， 手作業で設計したアプローチをより深いネットワークに一般化することは困難である。 同様に， 効率的なコーディングを 1 層以上に拡張することでいくつかの進歩があったが (26)， これらのアプローチでは効果的な深層ネットワークは得られなかった。 多層の HMAX ネットワークは， 既知の生物学的制約にほぼ一致するようにパラメータを選択することで作成された (12,13)。 HMAX ネットワークは， 下側頭 (IT) 皮質ニューロンの許容範囲 (12,27) や， 単一ユニットの選択性と許容範囲の間のトレードオフ (28) など， 高レベルの経験的観察を再現することに成功した。

しかし， 2000年代半ばになると， これらのアプローチはいずれも， V4 や IT などの高次皮質領域への拡張が困難であることが明らかになった。 例えば， HMAX モデルは， 視覚イメージのバッテリーにおける IT 集団の活動パターンと一致せず (29)， また， V4 と IT の神経データに適合した多層ニューラルネットワークは， 訓練データを過剰に適合させ，新規テスト画像では比較的小さな説明分散しか予測できなかった (8)。

成功しなかった理由の 1 つとして， 検討していた主にフィードフォワードのニューラルネットワークでは， データを効率的に取り込むには限界があったことが考えられる。 おそらく，フィードバック (30) やミリ秒単位のスパイクタイミング (31) を用いた， より洗練されたネットワークアーキテクチャが必要になるだろう。 2 つ目の可能性は， モデルのパラメータを適合させるのに十分な神経データがなかったために失敗したというものである。 単一ユニットの生理学的アプローチ (8) や全脳機能 MRI (29) では， おそらく 1,000 個の独立した刺激に対する反応を測定できるが， 配列電気生理学 (32) では， 約 10,000 個の刺激に対する反応を得ることができる。 今にして思えば， このようなネットワークを制約するために利用できる神経データの量は， 数桁も少なすぎた。

7. 新たな前進：ニューラルモデルとしてのゴール駆動型ネットワーク

ゴール駆動形なアプローチは， どのようなパラメータを使用しても， ニューラルネットワークが与えられた感覚システムの正しいモデルであるためには， 感覚システムがサポートする行動課題を解決するのに有効でなければならないという考えに基づいている。 このアプローチの考え方は， まず倫理的に適切な課題での成績のためにネットワークのパラメータを最適化し， ネットワークのパラメータが固定されたら， ネットワークとニューラルデータを比較するというものである。 このアプローチでは，純粋なニューラルフィッティングの深刻なデータ制限を回避することができる。 例えば， 物体認識の多くの難しい実例を含む何百万もの人間のラベル付き画像を収集することは， 同等のニューラルデータを得るよりもはるかに簡単である。 重要な問題は， このようなトップダウンの目標は， 生物学的構造を強く制約するのかということである。 ネットワークの出力で成績の最適化を行っても， ネットワークの隠れ層が， 例えば V1, V4, IT などの本物のニューロンのように振る舞うことができるのだろうか？ 最近の一連の結果は， 実際にそうなるかもしれないことを示している。

ゴール駆動形なアプローチの技術的基盤は， 人工知能課題のためにニューラルネットワークの性能を最適化するための近年の改良にある。 本節では， これらのツールがどのようにしてより優れたニューラルモデルを生み出したかを説明し， 次節では， これらのツールの基盤となる技術革新について説明する。

7.1 カテゴリー化に最適化された HCNN の最上位の隠れ層が IT ニューロンの反応を予測

何千もの HCNN モデルを， 課題成績と神経予測性の指標で評価したハイスループットな計算実験により， 重要な相関関係が明らかになった。 つまり， 高度な物体認識課題で優れた成績を発揮するアーキテクチャは， 皮質のスパイクデータをよりよく予測するということだ (33,34) (図2a)。 この考えをさらに推し進めるために， 最近の機械学習の進歩を利用して， 難しい物体分類課題で人間に近いレベルの性能を達成した階層型ニューラルネットワークモデルを発見した。 これらのモデルの最上位の隠れ層は， 腹側階層の最上位領域である IT 皮質のスパイク反応の定量的に正確な画像計算可能なモデルであることが判明した (18,33,34)(図2b,c)。 同様のモデルは， ヒト IT の機能的MRI データにおける集団の集合反応を予測することも示されている (図2d)(35,36)。

これらの結果は， 単に物体のカテゴリー同一性を反映した信号が IT 反応を予測できるというだけでは， 些細なこととして説明できない。 実際， 単一ニューロンレベルでは，IT 神経応答はほとんどカテゴリー的ではなく， カテゴリーとアイデンティティの情報に完全にアクセスできる理想的な観察者モデルは， ゴール駆動型 HCNN よりもはるかに精度の低い IT モデルとなる (33) (図2a,c)。 高いレベルの神経予測能力を得るためには， 真のイメージ計算可能なニューラルネットワークモデルであることが重要であると思われる。 言い換えれば，2 つの一般的な生物学的制約 (物体認識課題の行動的制約と HCNN モデルクラスによって課されるアーキテクチャ的制約) を組み合わせることで， 視覚感覚カスケードの複数の層のモデルを大幅に改善することができる。

これらのゴール駆動型モデルの最上位の隠れ層は， 最終的に IT 皮質のデータを予測するようになっているが，そうなるように明示的に調整されているわけではなく， 訓練過程で神経データに触れることは一切なかった。 モデルは 2 つの方法で一般化に成功した。 1 つ目は， ある意味的なカテゴリーに分類された実世界の写真を使ってカテゴリー認識の訓練を行ったが， 訓練で使った意味的なカテゴリーとは全く重ならない物体を含む合成された画像を使って， ニューロンとのテストを行ったことである。 第二に， ネットワークの学習に用いられた目的関数は， 神経データに適合するものではなく， 下流の行動目標 (例えば， カテゴリー化) に適合するものであった。 モデルのパラメータは， カテゴリー分類の成績を最適化するように独自に選択され， 非線形モデル層などの中間パラメータがすべて確定した後に，神経データと比較された。

別の言い方をすれば， HCNN のクラスの中には， 変動の多様な物体のカテゴリー化課題に対して， 質的に異なる， 効率的に学習可能な解が比較的少ないように見え， おそらく脳は， 進化と発達の時間軸の中で， そのような解を選ぶことを余儀なくされているのではないかと考えられる。 この仮説を検証するためには， カテゴリー化のために最適化されたときに高い性能を発揮する非 HCNN モデルを特定することが有用である。 この仮説は， そのようなモデルがあれば， 神経応答データを予測できないだろうと予測している。

8. 中間層と下層が V4 と V1 の反応を予測

IT にマッピングされる上位モデル層に加えて， 同じ HCNN モデルの中間層が， IT への主な皮質入力である中間視覚野， V4 皮質の神経応答の最先端の予測因子であることがわかった (33) (図2c)。 IT 皮質への適合は， モデルの最上位の隠れた層でピークに達するのに対し， V4 への適合は中間の層でピークに達する。 実際， これらの「偶然の」V4 に似た層は， その領域が何をしているかについての古典的な直観 (例えば， エッジ接続や曲率表現 (37)) から構築されたモデルよりも， V4 の反応を有意に予測している。 この傾向を引き継ぐように， ゴール駆動型 HCNN モデルの最下層は， ガボールウェーブレット のような活性化パターンを自然に含んでいる。 さらに， これらの最下層は， V1-V3 ボクセルデータの ボクセル応答の効果的なモデルを提供する (図2e) (35,36)。 このように，トップダウンの制約は， 腹側の階層にまで及んでいる。

視覚神経科学では， 下位の皮質領域におけるチューニングカーブ (例えば， V2 のエッジ結合 (文献38) や V4 の曲率 (文献39) など) を理解することが， 上位の視覚領域を説明するために必要な前提条件であると考えられている。 ゴール駆動型の深層 HCNN を用いた結果は， ボトムアップのプリミティブが特定されていなくても， トップダウンの制約によって中間領域の定量的に正確なモデルが得られることを示している (Box 3)。

Box 3

複雑な感覚システムにおける「理解」の意味について 複雑な神経系(67)を理解するとはどういうことなのか？ このパースペクティブでは， 成功するモデルは， 画像計算可能で， マッピング可能で， 定量的な予測が可能であることを示唆してきた。 しかし， これらの基準を満たすモデルは， 必ずしも理解を表しているのだろうか？ 深層ニューラルネットワークはブラックボックスであり， 説明しようとしている神経システムに対して限られた概念的洞察しか与えないと主張することができる。 実際， 深層 HCNN が非常に非線形な課題を実行する高度に複雑なシステムの内部応答を予測できるという事実そのものが， 初期のおもちゃモデルとは異なり， これらの深層モデルは初期のモデルよりも分析が困難であることを示唆している。 モデルの正しさと理解しやすさの間には，自然なトレードオフがあるのかもしれない。

最適な刺激と摂動の解析 しかし， 画像計算可能なモデルの重要な利点の 1 つは， 低コストで詳細な解析を行うことができ， ハイスループットの「仮想電気生理」を可能にすることである。 ターゲット画像の統計量に合わせて入力を最適化するか， 単一の出力ユニットの活性化を最大化するかのいずれかを行う最近の技術は， テクスチャ生成， 画像スタイルのマッチング， 最適な刺激合成において印象的な結果をもたらしている (文献 68 および Mordvintsev, A., Tyka, M. & Olah, C., http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html, 2015)。 これらの技術は， モデルのスケール効率を利用して， 巨大な刺激空間を現実的な実験手順を用いて測定できるほど小さなセットに縮小することで， 個々のニューロンの特徴的なドライバを特定するために使用することができる(69)。 また， 因果的介入実験 (70) にヒントを得て， モデル内のユニットに摂動を与えることで， 神経反応と行動の間の因果関係を予測し， 最も効果的な行動変化を得るために刺激や摂動パターンを最適化することもできる。

Marrの分析レベルを超えた具体的な例 ゴール駆動型のモデルは， より高いレベルの洞察をもたらす。 機能的な制約が神経的に予測可能なモデルを生み出すことは， 効率的符号化仮説 (23,24) を含む以前の研究を思い起こさせる。 どちらのアプローチでも， 最適化のための目的関数として表現される駆動概念が，パラメータがなぜそうなっているかを説明する。 効率的符号化とは異なり， ゴール駆動型 HCNN は， 生態学的な関連性が不明な過疎性のような抽象的な概念ではなく， 生物が行うことが知られている行動から目的関数を導き出すものである。 この意味で， 今回の研究は， システムの計算レベルの目標がアルゴリズムや実装レベルのメカニズムにどのように影響するかを調査する， Marr の分析レベル (71) に精神的に近いものである。 このアプローチは， 生物の自然な行動を研究して， その根底にある神経メカニズムについての洞察を得る神経倫理学にも関連している (72)。

8.1 皮質領域の生成モデルとしての HCNN 層

経験的に測定された各ニューロンに対して単一の非線形モデルを当てはめ， そこから得られたパラメータの分布を記述する従来のモデリングアプローチとは異なり(6)， 成績ベースのアプローチでは， すべてのニューロンに対して同時に単一のモデルを生成する。 その結果， 深層 HCNN の各層は， 対応する皮質領域の生成モデルとなり， そこから大量の (例えば) IT-, V4-, V1- のようなユニットを抽出することができる。 モデルの正しさを評価するために使用されたニューロンは， ランダムな電極サンプリングによって選ばれたものであるため， モデルのパラメータを更新したり， 新しい非線形関数を学習したりしなくても， 同じ領域からサンプリングされた将来のニューロンも同様によく予測されると考えられる。

8.2 聴覚野への応用

ゴールベースの HCNN モデルを， 視覚に比べて理解が進んでいない感覚領域に適用することは， 自然なアイデアである。 最も明白な候補は聴覚である。 音声認識， 話者識別，自然音識別 などの聴覚課題を解決するために最適化された HCNN モデルを作成することになる。 このようなモデルの中間層では， これまで知られていなかった非一次聴覚野の構造が明らかになるかもしれないという興味深い可能性がある。 初期の結果では，このアプローチが期待されている(40)。

9. HCNNの改良につながる要因

HCNN は， 神経科学からヒントを得て， 機械学習の中核的な道具となった。 HCNN は， 画像分類、顔識別， 位置特定， 行動認識， 深さ推定， その他のさまざまな視覚課題など， 多くの課題で成功を収めている (41)。 また， 関連するリカレント版のディープニューラルネットワークは， 音声認識の分野でも躍進している。 ここでは， このような最近の進歩をもたらした技術的な進歩について説明する。

10. フィルタのパラメータを最適化するためのハードウェアアクセラレーションによる確率的誤差逆伝播法

教師付き学習では， ある課題 (画像中の自動車の検出など) について， サンプル入力 (自動車と非自動車の画像など) と，各入力に対して望ましい結果を示すラベル (「自動車」や「犬」などの画像カテゴリラベルなど) を含む学習データセットを選択する。 そして， 学習アルゴリズムを用いて， ネットワークのパラメータ設定を最適化し， 出力層が学習データに望ましいラベルを与えるようにする (14)。 教師付きデータからフィルタのパラメータを学習するための強力なアルゴリズムとしては，バックプロパゲーションによる誤差勾配降下法 (14,42) が数十年前から存在する (Box4)。 しかし， 最近まで， バックプロパゲーションは， 大規模なデータセットでは計算が困難であった。 バックプロパゲーションの計算は， 単純な点演算や並列行列の点積が主なものであるため， 最近の GPU (Graphical Processing Unit) による高速プログラミングの登場は， 大きな恩恵をもたらした (15,33,43)。 フォン・ノイマン型CPUアーキテクチャよりもニューロモーフィックな GPUは， これらの演算に特に適しており， 日常的に 10 倍以上の速度向上を実現している (15)。 ニューロモーフィック・コンピューティングがさらに進化すれば， この傾向はさらに加速するだろう (44)。

Box 4 勾配逆伝播法

勾配逆伝播アルゴリズムの基本的な考え方は簡潔である。

1. 対象となる課題を，最小化すべき損失関数として定式化する (例: カテゴリー化誤差) 損失関数は， 入力 (画像など) とモデルパラメータの両方に対して区分的に微分可能でなければならない。
2. モデルのパラメータをランダムに， あるいは十分な情報に基づいた初期推測によって初期化する (14)。
3. 入力された学習サンプルごとに， フィルタのパラメータに対する誤差関数の微分を計算し， 入力データの合計値を算出する
4. ットワークのパラメータを勾配降下法で更新する。 つまり， 各パラメータを誤差の勾配と反対の方向に少しずつ動かしていく
5. 学習誤差が収束するか，オーバーフィットが懸念される場合には， 何らかの「早期停止」基準が満たされるまで， ステップ 3 と 4 を繰り返す(14)

フィードフォワードネットワークでこの手順を比較的効率的に行うことができるのは， 基本的な微積分の連鎖法則を適用するだけで， ある層のフィルター値に対する誤差の導関数を， すぐ上の層のものから効率的に計算することができるからである(42)。 微分の計算は， 最上層から始まり， ネットワークを逆にたどって最初の層に伝わっていく。 大規模なバックプロパゲーションを可能にしたもう一つの重要な技術革新は， 確率的勾配降下法 (SGD) である (42)。 SGD では， 学習データをランダムに選んだ小さなバッチに分割する。 勾配降下法は， 訓練データがなくなるまで， 各バッチに対して順に実行される。 その時点で， 通常は新たに選択されたランダムなバッチに対して手順を再開する。 SGD は， これまで考えられていたよりもはるかに大きなデータセットでのバックプロパゲーションを可能にし， 通常は安定した解に収束するが， そのような収束を保証する統計理論は十分に開発されていない。

11. アーキテクチャ・パラメータの自動学習手順

離散的なアーキテクチャパラメータ (例えば層数) は， 誤差バックプロパゲーションでは容易に最適化できない。 しかし， 離散的なパラメータは，最終的なネットワークの成績に重要である (15,18)。 従来， これらのパラメータは手作業で選択され， 改善が見られるまで様々な組み合わせを経験的にテストしていた。 最近では， ガウス過程最適化や遺伝的アルゴリズムなどの手法を用いて， より良いアーキテクチャパラメータを自動的に学習することができるようになった (15,45,46)。

12. ウェブ対応の大規模なラベル付きデータセット

最近の進歩のもう一つの重要な要因は， 大規模なラベル付きデータセットの登場である。 視覚領域では， 初期のデータセットは， 何百ものカテゴリの何百もの画像で構成されていることが多かった (47)。 しかし， このようなデータセットは， 計算機アーキテクチャを制約するのに十分な訓練データを提供するのに十分な規模と種類ではないことがわかった (15,48)。 このデータセットは，インターネット上のクラウドソーシングによって集められた， 何千ものカテゴリーの何千万もの画像を含んでいる (49)。 このような大規模なデータセットを活用するには， 上述のような効率的なハードウェアアクセラレーションによるアルゴリズムが必要であった。 このアルゴリズムを導入すれば， より深いニューラルネットワークを学習することができる。 バックプロパゲーションの学習サンプル数は， ネットワークのパラメータ数の 10 倍が目安とされている。 最新のディープネットワークのパラメータ数が 10 万をはるかに超えることを考えると， 少なくとも現在のパラメータ学習戦略では， 数百万の学習サンプルが必要であることが明らかである。 (脳で使用されているニューラル学習アルゴリズムは， ラベル付きデータを使用した場合， 現在の HCNN を学習するための計算方法よりも大幅に効率的であると思われ， 「10x」ヒューリスティックの対象にはならないかもしれない)

13. アーキテクチャークラスと訓練方法の小さな微調整の相乗効果

ニューラルネットワークのアーキテクチャーと訓練における他の多くの小さな変更が成績の向上に役立った。 特に重要な変更点は， 連続的に微分可能なシグモイド活性化関数を， 半正則化されたしきい値に置き換えることである (43)。 この活性化関数は， ほぼすべての場所で微分が一定またはゼロであるため，初期層の誤差勾配が小さすぎて効果的な最適化ができなくなる， いわゆる消失勾配問題の影響を受けにくくなる。 2 つ目の改良点は， バックプロパゲーションの際にネットワークにノイズを注入して， 脆弱でオーバーフィットな重みパターンの学習を防ぐ正則化法の導入である (43)。

14. エンジニアリングの理不尽な有効性

最近の改良は， いくつかの重要な工学的改良の積み重ねである (例えば，文献 50,51)。 これらの変更は， 数十年前に記述されたオリジナルの HCNN やバックプロパゲーションの概念を超える大きな概念的ブレークスルーを示唆するものではないかもしれないが， それにもかかわらず， 最終的な結果の膨大な改善につながった。 大規模なデータセットと慎重なエンジニアリングは， 当初の予想よりもはるかに重要であった (52)。

15. 今後の展開：可能性と限界

ゴール駆動型ディープニューラルネットワークモデルは， 3 つの基本要素から構築される(図3)。

- 脳の解剖学的および機能的結合に関する知識を形式化した， システムを構築するためのモデルアーキテクチャクラス。
- システムが達成すべき行動目標 (物体の分類など)， および
- 行動目標を達成するために， モデルクラス内のパラメータを最適化する学習ルール

以上の結果は， これら 3 つの要素を組み合わせることで， これまでの感覚皮質モデルを大幅に上回る， 神経データを検証可能な形で予測する詳細な計算モデルを構築できることを示している。 今後の進歩は， これら 3 つの要素とその限界をよりよく理解することにかかっている (Box 5)。

> Box 5
>
> **敵対的最適化効果の理解**
>
> HCNN 研究における最近の興味深い進展は， 敵対的画像の発見である。 これは， 通常の写真に人間には検出できないような微妙な修正を加え， ネットワークが修正された画像内の任意のオブジェクトを誤って検出するようにしたものである (73,74)。 つまり， 敵対的画像は， 既存の HCNN が， 人間を騙すのとは質的に異なる種類の錯覚に陥りやすいことを示しているのである。 この画像は， 敵対的最適化によって作成されている。 敵対的最適化とは， ネットワークの最終的なカテゴリー検出層に最大の変化をもたらすように， 原画像の画素を最適に修正する処理過程である。 このような画像を作成するには， ネットワークの内部パラメータに完全にアクセスする必要があるが， 物理的な世界では自然に発生することはない。
>
> 前述のゴール駆動形モデリングの 3 つの要素に沿って考えると (図3参照)， 敵対的な例を説明するためのいくつかの可能性がある。 (例えば， ある人間が騙されても， 他の人間が正しく認識する特異な画像を作成することなどは， 実験が個々の脳の詳細な微小回路にアクセスし， その上で敵対的最適化アルゴリズムを実行することができれば， 同様の効果が人間にも再現可能であることを示している。 (カテゴリー化目標に対する最適化は脆弱だが， より豊かで頑健な最適化目標を用いれば， その効果はなくなるだろう。 逆説的な例は， 脳モデルとしての HCNN の基本的なアーキテクチャ上の欠陥を露呈しており， 他のネットワーク構造(例えば， 再帰) を組み込むことによってのみ， 逆説的な例を克服することができる，というものである。 どちらが正しいかは別にして，敵対的最適化の効果を理解することは， HCNN 自体， 特に脳の推定モデルとしてのHCNN をよりよく理解するための重要な要素であると思われる。

15. アーキテクチャクラスの改善

感覚皮質を理解するために計算モデルを継続的に使用するには， モデル層と皮質領域の間のマッピングをより詳細かつ明示的に行う必要がある。 テンプレートマッチングやプーリングなどの HCNN の操作は神経学的にはもっともらしいが， HCNN で使われているパラメータ化が実際に皮質の微小回路とどのようにつながっているかを理解することははるかに困難である。 同様に， HCNN のモデル層の階層は， 観察された腹側皮質領域の全体的な順序とおおむね一致しているように見えるが， モデル層と脳領域の一致が一対一 (あるいはそれに近い状態) であるかどうかは， 完全には理解されていない。 最近の高性能なコンピュータビジョンネットワークでは， 層の数が大幅に増え， 時には 20 以上になることもある(50)。 このような非常に深いネットワークが， 神経データをよりよく説明できるかどうかを評価することは重要である。 なぜなら， 神経の適合性からの逸脱は， アーキテクチャの選択が脳内のものとは異なることを示唆するからである。 より一般的には， HCNN のクラスの中で， カテゴリー化の成績に最適化された場合， どのアーキテクチャが腹側経路の神経反応データに最も適合するのか， という問いがある。 上記の結果は， これが成人の腹側経路のアーキテクチャを推論する新しい方法になりうることを主張している。

もちろん， このようなトップダウン型の性能重視のアプローチは， 二光子顕微鏡法， オプトジェネティクス， 電子顕微鏡再構成法， その他アーキテクチャのクラスをより直接的に絞り込むことを目的としたくんれん技術など， 最先端の実験技術と組み合わせるべきである。 神経回路レベルでの経験的な理解が深まれば， 生物学的に関連性のある HCNN のクラスを絞り込むことができ， 特定のアーキテクチャを除外したり， フィルタのパラメータについて情報に基づいた初期推測を行ったりすることができる。 そうすれば， モデルはより少ないパラメータを学習するだけで， 同等以上の神経予測能力を得ることができる。

視覚でも聴覚でも， 皮質下の構造に関する既知の結果を利用して， より生物学的に現実的な感覚フロントエンドを初期層に組み込むことで， モデルアーキテクチャクラスを向上させることができる (53)。 その反対に， 高次の皮質領域 (例えば， 顔のパッチ) には， 大規模な空間的不均一性がある (4)。 HCNN の低層では， 網膜地図を介して皮質の表面に明らかにマッピングされているが， 高層ではこの関係はあまり明確ではない。 多次元の深層ネットワークの出力が 2 次元の皮質シートにどのようにマッピングされるのか， また， このことが機能的組織化にどのような影響を与えるのかを理解することは， 重要な未解決問題である。

16. ゴールと訓練セットの理解を深める

目標と訓練セットの選択は, モデル開発に大きな影響を与えており, 高変動データセットは実世界のカテゴリーにおける真の異質性を明らかにしている (33,48,49)。 このデータ主導の傾向は今後も続くと思われる(52)。 最近の重要な結果は, ある課題 (例えば, ImageNet 分類) のために訓練された HCNN が, 最初に訓練された課題とは全く異なる他の多くの視覚課題に一般化することである (41)。 カテゴリー化に関連する多くの課題が「無料」でついてくるとしたら, どの課題はそうではないのだろうか？ 特に重要な課題は, カテゴリー化の最適化では解決できず, むしろ独立した直接の最適化を必要とする課題を見つけ出し, これらの課題に最適化されたモデルをテストして, 腹側経路の神経データをよりよく説明できるかどうかを確認することである。 この目標を達成するためには, 豊富な新しいラベル付きデータセットの開発が不可欠である。 様々な感覚課題に対応する HCNN システムが, アーキテクチャの共有や乖離の観点から, 互いにどのように関連しているのかを理解することは, 感覚ドメイン (54) 内だけでなく, ドメイン間 (例えば, 視覚と聴覚の間 :図3参照) でも興味深いことである。

17. 学習ルールの理解を深める

教師付き学習が実際の知覚システムに著しく適合した作業モデルを作成することは貴重だが, 大脳皮質が正確なバックプロパゲーションを実装していることは生理的にあり得ない。 現在の深層学習アプローチと実際の生物学的学習との間にある中核的な矛盾は, 効果的な HCNN を訓練するには, 非常に多くの高レベルの意味的ラベルが必要であることである。 人間や高等霊長類, その他の動物における真の生物学的な生後学習では, 大量の教師なしデータを使用することはあっても, 外部からのラベル付きの監視をこれほど大量に必要とすることはないだろう。 生物学的に現実的な教師なしまたは半教師付き学習アルゴリズム (55-57) を発見し, 高いレベルの性能と神経予測性を生み出すことができれば, 人工知能と神経科学の両方の観点から興味深いものとなるだろう。

18. 感覚システムとフィードフォワードネットワークを超えて

大規模なフィードフォワード HCNN は, ワーキングメモリを含む, 拡張可能な状態を保存する脳システムのダイナミクスを完全に説明することはできない。 なぜなら、フィードフォワードネットワークのダイナミクスは, 入力履歴とは無関係に同じ状態に収束するからである。 しかし, 注意, 意思決定, 運動プログラム生成などの神経現象にリカレント神経ネットワークを関連付ける文献が増えてきている (58)。 深層ニューラルネットワークでモデル化されるような豊かな感覚入力システムと, これらのリカレントネットワークを組み合わせたモデルは, 感覚モデルがしばしば投げかけられる純粋な「表現」の枠組みから抜け出して, 単純な分類や二値的な意思決定を超えた, より洗練された認知行動を探求するための実り多い道を提供する可能性がある。 これは, 行動出力と入力刺激の間に複雑なループがある場合, 例えば, 複雑な感覚環境での長い時間スケールでのエージェントの探索をモデル化する場合などに特に興味深い (59)。 強化学習 (60) から得られた最近の興味深い結果は, 戦略学習の問題を解決する上で, ディープニューラルネットワーク技術がいかに強力であるかを示している。 これらの結果を, 腹側視覚野と, 例えば頭頂葉皮質や海馬との間のインターフェースに関する神経科学のアイデアにマッピングすることは, 非常に興味深いことである (61,62)。

19. 結論

以上のように, 深層ニューラルネットワークは, 神経科学者が感覚システムの定量的に正確な計算モデルを作成する能力に変革をもたらしつつある。 このようなモデルは, すでにいくつかの注目すべき結果を出しており, ヒトとサルの両方における複数の神経科学データを説明している (33-36)。 しかし, 重要な科学的進歩と同様に, これらのアイデアは, 答えと同じくらい多くの新しい問題を提起しています。 今後も, 神経科学, コンピュータ科学, 認知科学の 3 つの分野が相互に協力し合いながら, 刺激的でやりがいのある研究を続けていく必要がある。