bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Why Does the Neocortex Have Layers and Columns, A Theory of Learning the 3D Structure of the World

Jeff Hawkins*, Subutai Ahmad, Yuwei Cui

Numenta, Inc,
Redwood City, California, USA

*Corresponding author
Emails: jhawkins@numenta.com, sahmad@numenta.com, ycui@numenta.com

A version of this manuscript has been submitted for publication. This is a draft preprint. The final paper will be different from this version.

We welcome comments. Please contact the authors with any questions or comments.

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Why Does the Neocortex Have Layers and Columns, A Theory of Learning the 3D Structure of the World

Jeff Hawkins, Subutai Ahmad, and Yuwei Cui
Numenta, Inc, Redwood City, California, United States of America

ABSTRACT

Neocortical regions are organized into columns and layers. Connections between layers run mostly perpendicular to the surface suggesting a columnar functional organization. Some layers have long-range lateral connections suggesting interactions between columns. Similar patterns of connectivity exist in all regions but their exact role remains a mystery. In this paper, we propose a network model composed of columns and layers that performs robust object learning and recognition. Columns integrate their changing inputs over time to learn complete models of observed objects. Lateral connections across columns allow the network to more rapidly infer objects based on the partial knowledge of adjacent columns. Because columns integrate input over time and space, the network learns models of complex objects that extend well beyond the receptive field of individual cells. Our network model introduces a new feature to cortical columns. We propose that a representation of location is calculated within the sub-granular layers of each column. The representation of location is relative to the object being sensed. Pairing sensory features with locations is a requirement for modeling objects and therefore must occur somewhere in the neocortex. We propose it occurs in every column in every region. Our network model contains two layers and one or more columns. Simulations show that small single-column networks can learn to recognize hundreds of complex multi-dimensional objects. Given the ubiquity of columnar and laminar connectivity patterns throughout the neocortex, we propose that columns and regions have more powerful recognition and modeling capabilities than previously assumed.

INTRODUCTION

The neocortex is complex. Within its 2.5mm thickness are dozens of cell types, numerous layers, and intricate connectivity patterns. The connections between cells suggest a columnar flow of information across layers as well as a laminar flow within some layers. Fortunately, this complex circuitry is remarkably preserved in all regions, suggesting that a canonical circuit consisting of columns and layers underlies everything the neocortex does. Understanding the function of the canonical circuit is a key goal of neuroscience.

Over the past century, several theories have been proposed to explain the existence of cortical layers and columns. One theory suggested these anatomical constructs minimize the amount of wiring in cortical tissues (Shipp et al., 2007). Some researchers suggested there should be functional differentiation of different cortical layers that match the anatomical structure (Douglas and Martin, 2004). Others have proposed that long-range laminar connections contribute to attention-related changes in receptive field properties (Raizada and Grossberg, 2003). Recent advances in recording technologies now enable detailed recording of activity in the micro-circuitry of cortical columns. However, despite these advances, the function of networks of neurons organized in layers and columns remains unclear. Due to lack of analytic insight, assigning any function to columns remains controversial (Horton and Adams, 2005).

Lacking a theory of why the neocortex is organized in columns and layers, almost all artificial neural networks, such as those used in deep learning (LeCun et al., 2015) and spiking neural networks (Maass, 1997), do not include these features, introducing the possibility they may be missing key functional aspects of biological neural tissue. To build systems that work on the same principles as the neocortex we need an understanding of the functional role of columnar and laminar projections.

Cellular layers vary in the connections they make, but a few general rules have been observed. Cells in layers that receive direct feedforward input do not send their axons outside the local region and they do not form long distance horizontal connections within their own layer. Cells in layers that are driven by input layers form long range connections within their layer, and also send an axonal branch outside of the region, constituting an output of the region. This two-layer input-output circuit is a persistent feature of cortical regions. The most commonly recognized instance involves layer 4 and Layer 2/3. Layer 4 receives feedforward input. It projects to layer 2/3 which is an output layer. Upper layer 6 also receives feedforward input. It projects to layer 5, which is an output layer, and therefore layers 6 and 5 may be a second instance of the two-layer input-output circuit. The prevalence of this two-layer connection motif suggests it plays an essential role in cortical processing.

In this paper, we introduce a theory of how columns and layers learn the structure of objects in the world. It is a sensorimotor theory in that learning and inference require movement of sensors relative to objects. We also introduce a network model based on the theory. The network consists of one or more columns, where each column contains an input layer and an output layer. First, we show how even a single column can learn the structure of complex objects. A single column can only sense a part of an object at any point in time, however, the column will be exposed to multiple parts of an object as the corresponding sensory organ moves. While the activation in the input layer changes with each movement of the sensor, the activation in the output layer remains stable, associating a single output representation with a set of feature representations in the input layer. Thus, a single cortical column can learn models of complete objects through movement. These objects can be far larger than any individual cell's receptive field.

Next, we show how multiple columns collaborate via long-range intralaminar connections. At any point in time, each column has only partial knowledge of the object it is observing, yet adjacent columns are typically sensing the same object, albeit at different locations on the object. Long range connections in the output layer allow multiple columns to rapidly reach a consensus of what object is being

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

observed. Although learning always requires multiple sensations via movement, inference can often occur in a single or just a few sensations. Through simulation we illustrate that our model can learn the structure of complex objects, it learns quickly, and it has high capacity.

A key component of our theory is the presence in each column of a signal representing location. The location signal represents an "allocentric" location, meaning it is a location relative to the object being sensed. In our theory, the input layer receives both a sensory signal and the location signal. Thus, the input layer knows both what feature it is sensing and where the sensory feature is on the object being sensed. The output layer learns complete models of objects as a set of features at locations. This is analogous to how computer-aided-design programs represent multi-dimensional objects.

Because different parts of a sensory array (for example different fingers or different parts of the retina) sense different parts of an object, the location signal must be calculated uniquely for each sensory patch and corresponding area of neocortex. We propose that the location signal is calculated in the sub-granular layers of cortical columns and is passed to input layer 4 via projections from layer 6.

It is important to note that we deduced the existence of the allocentric location signal. We first deduced its presence by considering how fingers can predict what they will sense while moving and touching an object. However, we believe the location signal is present in all neocortical regions. We show empirical evidence in support of this hypothesis. Although we cannot yet propose a complete mechanism for how the location signal is derived, the task of determining location and predicting new locations based on movement is similar to what grid cells do in the medial entorhinal cortex. Grid cells offer an existence proof that predictive models of allocentric location are possible, and they suggest mechanisms for how the location signal might be derived in cortical columns.

The theory is consistent with a large body of anatomical and physiological evidence. We discuss this support and propose several predictions that can be used to further test the theory.

MODEL

Motivation

Our research is focused on how networks of neurons in the neocortex learn predictive models of the world. Previously, we introduced a network (Hawkins and Ahmad, 2016) that learns a predictive model of naturally changing sensory sequences. In the present paper, we extend this network to address the related question of how the neocortex learns a predictive model of static objects, where the sensory input changes due to our own movement.

A simple thought experiment may be useful to understand our model. Imagine you reach your hand into a black box and try to determine what object is in the box, say a coffee cup. Using only one finger it is unlikely you could identify the object with a single touch. However, after making one contact with the cup, you move your finger and touch another location, and then another. After a few touches, you identify the object as a coffee cup. Recognizing the cup requires more than just the tactile sensation from the finger, the brain must also integrate knowledge of how the finger is moving, and hence where it is relative to the cup. Once you recognize the cup, each additional movement of the finger generates a prediction of where the finger will be on the cup after the movement, and what the finger will feel when it arrives at the new location. This is the first problem we wanted to address, how a small sensory array (e.g. the tip of a finger) can learn a predictive model of three dimensional objects by integrating sensation and movement-derived location information.

If you use two fingers at a time you can identify the cup with fewer movements. If you use five fingers you will often be able to identify an object with a single grasp. This is the second problem we wanted to address, how a set of sensory arrays (e.g. tips of multiple fingers) work together to recognize an object faster than they can individually.

Somatic inference is obviously a sensorimotor problem. However, vision and audition are also sensorimotor tasks. Therefore, the mechanisms underlying sensorimotor learning and inference should exist in all sensory regions, and any proposed network model should map to the detailed anatomical and physiological properties that exist in all cortical regions. This mapping, an explanation of common cortical circuitry, is a third goal of our model.

Model description

Our model extends previous work showing how a single layer of pyramidal neurons can learn sequences and make predictions (Hawkins and Ahmad, 2016). The current model consists of two layers of pyramidal neurons arranged in a column. The model has one or more of these columns (**Figure. 1A**). Each cortical column processes a subset of the sensory input space and is exposed to different parts of the world as the sensors move. The goal is to have the output layer of each column converge on an object representation that is consistent with the accumulated sensations over time and across all columns.

The input layer of each column in our model receives a sensory input and a location input. The sensory input is a sparse binary array representing the current feature in its input space. The location input is a sparse binary array representing the location of the feature on the object. There are numerous observations in the neocortex that receptive fields are modified by location information. Grid cells in the entorhinal cortex also solve a similar location encoding problem and therefore represent a model of how location might be derived in the neocortex. We explore these ideas further in the discussion section. For our model we require a) that the location of a feature on an object is independent of the orientation of the object, and b) that nearby locations have similar representations. The first property allows the system to make accurate predictions when the object is sensed in novel positions relative to the body. The second property enables noise tolerance – you don't have to always sense the object in precisely the same locations.

Below we describe our neuron model, the connectivity of layers and columns, and how the sensory and location inputs are combined over time to recognize objects. A more detailed description of the activation and learning rules is available in the Methods section.

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

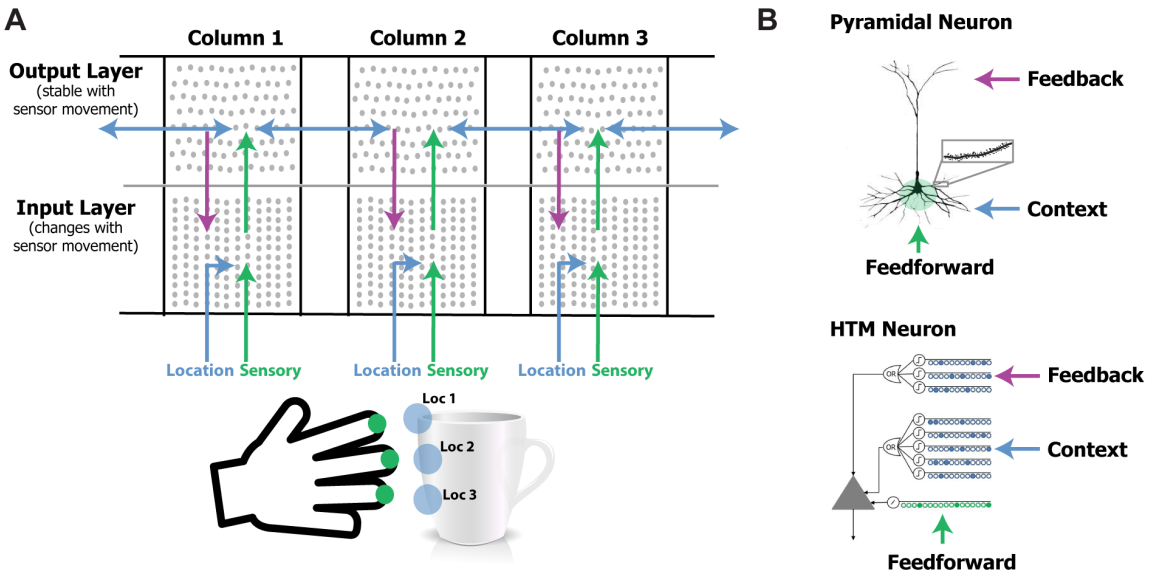

Figure 1 A. Our network model contains one or more laterally connected cortical columns (three shown). Each column receives feedforward sensory input from a different sensor array (e.g. different fingers or adjacent areas of the retina (not shown)). The input layer combines sensory input with a modulatory location input to form sparse representations that correspond to features at specific locations on the object. The output layer receives feedforward inputs from the input layer and converges to a stable pattern representing the object (e.g. a coffee cup). Convergence in the second layer is achieved via two means. One is by integration over time as the sensor moves relative to the object, and the other is via modulatory lateral connections between columns that are simultaneously sensing different locations on the same object (blue arrows in upper layer). Feedback from the output layer to the input layer allows the input layer to predict what feature will be present after the next movement of the sensor. **B**. Pyramidal neurons have three synaptic integration zones, proximal (green), basal (blue), and apical (purple). Although individual synaptic inputs onto basal and apical dendrites have a small impact on the soma, co-activation of a small number of synapses on a dendritic segment can trigger a dendritic spike (*top right*). The HTM neuron model incorporates active dendrites and multiple synaptic integration zones (*bottom*). Patterns recognized on proximal dendrites generate spikes. Patterns recognized on the basal and apical dendrites depolarize the soma without generating a spike. Depolarization is a predictive state of the neuron. Our network model relies on these properties and our simulations use HTM neurons.

Neuron model: We use HTM model neurons in the network (Hawkins and Ahmad, 2016). HTM neurons incorporate dendritic properties of pyramidal cells (Spruston, 2008), where proximal, basal, and apical dendritic segments have different functions (**Figure 1B**). Patterns detected on proximal dendrites represent feedforward driving input, and can cause the cell to become active. Patterns recognized on a neuron's basal and apical dendrites represent modulatory input, and will depolarize the cell without immediate activation. Depolarized cells fire sooner than, and thereby inhibit, non-depolarized cells that recognize the same feedforward patterns. In the rest of the paper we refer to proximal dendritic inputs as feedforward inputs, and the distal basal and apical dendritic inputs as modulatory inputs. A detailed description of functions of different dendritic integration zones can be found in (Hawkins and Ahmad, 2016).

Input layer: The input layer consists of HTM neurons arranged in minicolumns. (Here a minicolumn denotes a thin vertical arrangement of neurons (Buxhoeveden, 2002).) For our simulations we typically have 100-300 minicolumns per cortical column. The feedforward input of cells in this layer is the sensory input. As in (Hawkins and Ahmad, 2016) cells within a minicolumn recognize the same feedforward patterns. We map each sensory feature to a sparse set of minicolumns,

The modulatory input for cells in the input layer represents the location on an object. During learning, one cell in each active minicolumn is chosen to learn the current modulatory location signal. During inference, only cells that recognize both the modulatory location input and the feedforward driving input will become active. In this way, the input layer forms a sparse representation that is unique for a specific sensory feature at a specific location on the object.

Output layer: The output layer also contains HTM neurons. The set of active cells in the output layer represents objects. Cells in the output layer receive feedforward driver input from the input layer. During learning, the set of cells representing an object remain active over multiple movements and learn to recognize successive patterns in the input layer. Thus, an object comprises a representation in the output layer, plus an associated set of feature/location representations in the input layer.

The modulatory input of cells in the output layer comes from other output cells representing the same object, both from within the column as well as from neighboring columns via long-range lateral connections. As in the input layer, the modulatory input acts as a bias, and cells representing the same object will positively bias each other. Cells with more modulatory input will win and inhibit cells with less modulatory input. Thus, if a column has feedforward support for objects A and B at time t, and feedforward support for objects B and C at time t+1, the output layer will converge onto the representation for object B at time t+1 due to modulatory input from time t. Similarly, if column 1 has feedforward support for objects A and B, and column 2 has feedforward support for objects B and C, the output layer in both columns will converge onto the representation for object B.

Feedback connections: Neurons in the input layer receive

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

feedback connections from the output layer. Feedback input representing an object, combined with modulatory input representing an anticipated new location due to movement, allows the input layer to predict the next sensory input. In our model, feedback is an optional component. If included, it improves robustness to sensory noise and ambiguity of location.

Illustrative example

Figure 2 illustrates how the two layers of a single column cooperate to disambiguate objects that have shared features, in this case a cube and a wedge. The first sensed feature-location, labeled $\mathbf{f_1}$, is ambiguous, as it could be part of either object. Therefore, the output layer simultaneously invokes a union of representations, one for each object that has that feature at that location. Feedback from the output layer to the input layer puts cells in a predictive state (shown in red). The predicted cells represent the set of all feature-locations consistent with the set of objects active in the output layer. The red cells thus represent the predictions of the network consistent with the sensations up to this point. Upon the second sensation, labeled $\mathbf{f_2}$, only the subset of cells that is consistent with these predictions and the new feature become active. Each subsequent sensation narrows down the set until only a single object is represented in the output layer.

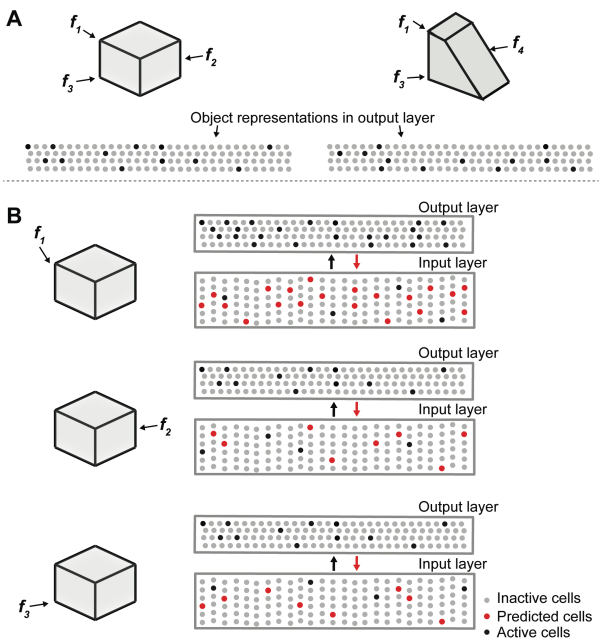

Figure 2 Cellular activations in the input and output layers of a single column during a sequence of touches on an object. **A.** Two objects (cube and wedge) and their associated output layer representations. For each object, three feature-location pairs are shown. $\mathbf{f_1}$ and $\mathbf{f_3}$ are common to both the cube and wedge. **B**. Cellular activations in both layers caused by a sequence of three touches on the cube (in time order from top to bottom). The first touch (at $\mathbf{f_1}$) results in a set of active cells in the input layer (black dots in input layer) corresponding to that feature-location pair. Cells in the output layer receive this representation through their feed-forward connections (black arrow). Since the input is ambiguous, the output layer forms a representation that is the union of both the cube and the wedge (black dots in output layer). Feedback from the output layer to the input layer (red arrow) causes all feature-location pairs associated with both potential objects to become predicted (red dots). The second touch (at $\mathbf{f_2}$) results in a new set of active cells in the input layer. Since $\mathbf{f_2}$ is not shared with the wedge, the representation in the output layer is reduced to only the cube. The set of predicted cells in the input layer is also reduced to the feature-location pairs of the cube. The third touch (at $\mathbf{f_3}$) would be ambiguous on its own, however, due to the past sequence of touches and self-reinforcement in the output layer, the representation of the object in the output layer remains unique to the cube. Note the number of cells shown is unrealistically small for illustration clarity.

Learning

The learning rule is based on simple Hebbian-style adaptation: when a cell fires, previously active synapses are strengthened and inactive ones are weakened. There are two key differences with most other neural models. First, learning occurs at the level of the dendritic segment, not the entire neuron (Stuart and Häusser, 2001; Losonczy et al., 2008). Second, the model neuron learns by growing and removing synapses from a pool of potential synapses (Chklovskii et al., 2004). A complete description of the biological motivation of the synaptic learning rules can be found in (Hawkins and Ahmad, 2016). Below we briefly describe how these principles enable the network to learn; the formal learning rules are described in the Materials and Methods section.

The input layer learns specific feature/location combinations. For a given input, if a feature/location combination has not been previously learned, no cell is predicted. A random cell from each minicolumn corresponding to that feature is chosen as the winner and becomes active. Each winner cell learns by forming modulatory connections with the current location input. If the location input is encountered again the corresponding set of cells will be depolarized. If the expected sensory feature arrives, the depolarized cells will fire first, and both the feedforward and modulatory inputs will be reinforced. Apical dendrites of the winning cells form connections to active cells in the output layer.

The output layer learns representations corresponding to objects. When the network first encounters a new object, a sparse set of cells in the output layer is randomly selected to represent the new object. These cells remain active while the system senses the object at different locations. Connections between the changing active cells in the input layer and unchanging active cells in the output layer are continuously reinforced. Thus, each output cell pools over multiple feature/location representations in the input layer. Dendritic segments on cells in the output layer form lateral modulatory connections to active cells within their own column, and to active cells in nearby columns. A reset occurs when the system switches to a new object, and a different set of cells will be selected to represent the new object.

SIMULATION RESULTS

In this section, we describe simulation results that illustrate the performance of our network model. The network structure consists of two layers, as described earlier (**Figure 1**). In the first set of simulations the input layer of each column consists of 150 minicolumns, with 16 cells per minicolumn. The output layer of each column consists of 4096 cells. The output layer contains inter-column and intra-column connections via the distal basal dendrites of each cell. The output layer also projects back to the apical dendrites of the input layer within

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the same column. All connections are continuously learned and adjusted during the training process.

We trained the network on a library of up to 500 objects (**Figure 3**). Each object consists of 10 sensory features chosen from a library of 5 to 30 possible features. Each feature is assigned a corresponding location on the object. Note that although each object consists of a unique set of features/locations, any given feature or feature/location is shared across several objects. As such, a single sensation by a single column is insufficient to unambiguously identify an object.

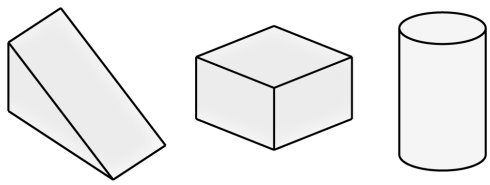

Figure 3 An illustrative subset of the 3D objects used for training. Note that individual local sensory features, such as a corner, appear on multiple objects, and often in the same location on different objects. Therefore, a single sensation is insufficient for recognition when using a single column. As the number of trained objects increases the number of shared feature/locations also increases, although the set of feature/locations for each object is unique.

The set of active cells in the output layer represents the objects that are recognized by the network. During training the representation for each object is chosen randomly. During inference we say that the network unambiguously recognizes an object when the representation of the output layer overlaps significantly with the representation for correct object and not for any other object. (Complete details of object construction and recognition are described in Materials and Methods).

In the following paragraphs we first describe network convergence, using single and multi-column networks. We then discuss the capacity of the network.

Network convergence

As discussed earlier, the representation in the output layer is consistent with the recent sequence of sensed features and locations. Multiple output representations will be active simultaneously if the sensed features and locations are not unique to one particular object. The output converges to a single object representation over time as the object is explored via movement. **Figure 4** illustrates the rate of convergence for a one column network and for a three-column network. Multiple columns working together reduces the number of sensations needed for recognition.

In **Figure 5A** we plot the mean number of sensations required to unambiguously recognize an object as a function of the total number of objects in the training set. As expected, the number of sensations required increases with the total number of stored objects. However, in all cases the network is able to eventually recognize objects uniquely. The number of sensations is also dependent on the overall confusion between the set of objects**.** The more unique the objects, the faster the network can disambiguate them.

Figure 5B illustrates the mean number of sensations required to recognize an object as a function of the number of cortical columns in the network. The number of sensations decreases rapidly as the number of columns increases. With six cortical columns, the network almost always recognizes the object with a single sensation. In this experiment each column receives lateral input from every other column.

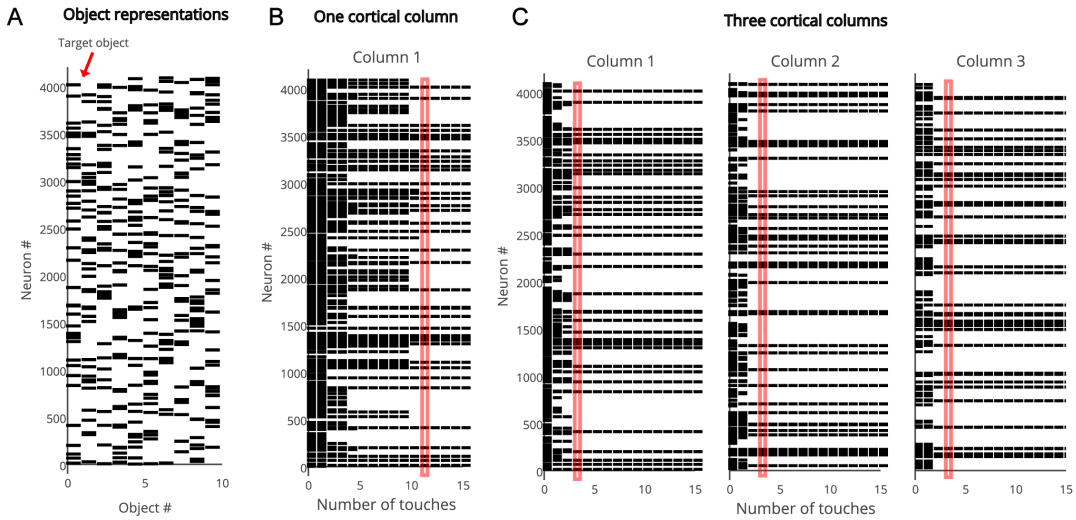

Figure 4. A. The output layer represents each object by a sparse pattern. We tested the network on the first object. **B**. Activity in the output layer of a single column network as it touches the object. The network converges after 11 touches (red rectangle). **C**. Activity in the output layer of a three column network as it touches the object. The network converges much faster, after 4 touches (red rectangle). In both B and C the representation in Column 1 is the same as the target object representation after convergence.

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

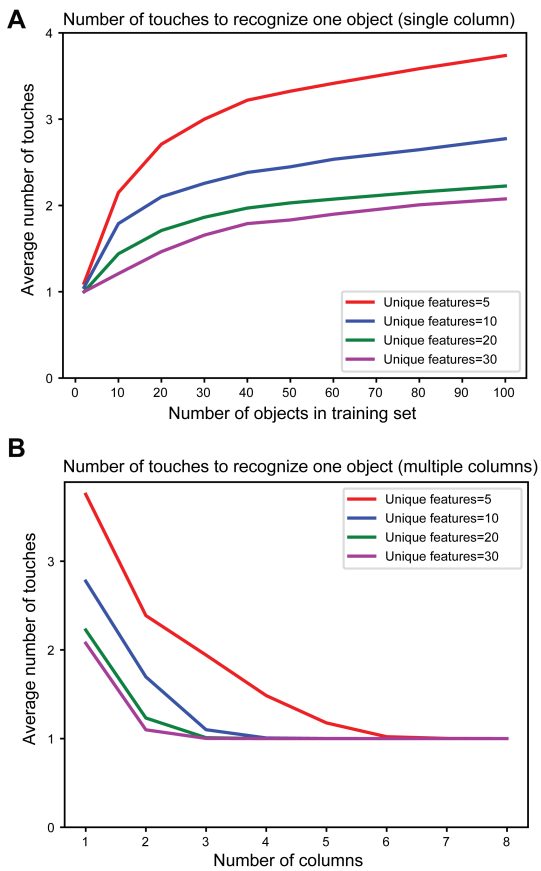

Figure 5 A. Mean number of observations to unambiguously recognize an object with a single column network as the set of learned objects increases. We train models on varying numbers of objects, from 1 to 100 and plot the average number of touches required to unambiguously recognize a single object. The different curves show how convergence varies with the total number of unique features from which objects are constructed. In all cases the network is able to eventually recognize the object, but the recognition is much faster when the set of features is greater. **B**. Mean number of observations to unambiguously recognize an object with multi-column networks as the set of learned objects increases. We train each network with 100 objects and plot the average number of touches required to unambiguously recognize an object. Recognition time improves rapidly as the number of columns increases.

Capacity

We define capacity as the maximum number of objects a network can learn and recognize without confusion. We analyzed four different factors that impact capacity: the representational space of the network, the number of minicolumns in the input layer, the number of neurons in the output layer, and the number of cortical columns. In our analysis we used numbers similar to those reported in experimental data. For example, cortical columns vary from 300 μm to 600 μm in diameter (Mountcastle, 1997), where the diameter of a minicolumn is estimated to be in the range of 30-60 μm (Buxhoeveden, 2002). For our analysis and simulations we assumed a cortical column contains between 150 and 250 minicolumns.

First, the neural representation must allow the input and output layers to represent a large number of unique feature/locations or objects, respectively. As illustrated in **Figure 2**, both layers use sparse representations. Sparse representations have several attractive mathematical properties that allow robust representation of a very large number of elements (Ahmad and Hawkins, 2016). With a network of 150 minicolumns, 16 cells per minicolumn, and 10 simultaneously active minicolumns, we can uniquely represent $\binom{150}{10}\sim 10^{15}$ sensory features. Each feature can be represented at 16^{10} different locations. Similarly, the output layer can represent $\binom{n}{w}$ unique objects, where n is the number of output cells and w is the number of active cells at any time. With such a large representational space, it is extremely unlikely for two feature/location pairs to have a significant number of overlapping bits by chance (Supplementary material). Therefore, it is safe to assume that the number of objects that can be uniquely represented in the output layer is not a limiting factor in the capacity of the network.

As the number of learned objects increases, neurons in the output layer form increasing numbers of connections to neurons in the input layer. If an output neuron connects to too many input neurons, it may be falsely activated by a pattern it was not trained on. Therefore, the capacity of the network is limited by the pooling capacity of the output layer. Mathematical analysis suggests that a single cortical column can store hundreds of objects before reaching this limit (see Supplementary material).

To measure actual network capacity we trained networks with an increasing number of objects and plotted retrieval accuracy. For a single cortical column, with 4,096 cells in the output layer and 150 minicolumns in the input layer, the retrieval accuracy remains perfect up to 400 objects (**Figure 6A**, blue). The retrieval accuracy drops when the number of learned objects exceeds the capacity of the network.

From the mathematical analysis, we expect the capacity of the network to increase as the size of the input and output layers increase. We again tested our analysis through simulations. With the number of active cells fixed, the capacity increases with the number of minicolumns in the input layer (**Figure 6A**). This is because with more cells in the input layer, the sparsity of activation increases, and it is less likely for an output cell to be falsely activated. The capacity also increases with the number of output cells when the size of the input layer is fixed (**Figure 6B**). This is because the number of feedforward connections per output cell decreases when there are more output cells available.

We find that networks with more cortical columns have higher capacity (**Figure 6C**). In this case the number of connections in each individual cortical column is independent of the number of columns. Nevertheless, the lateral connections in the output layer help to disambiguate inputs when single cortical columns hit their capacity limit. The impact of cortical column number on capacity is less dramatic than changing the size of the input and output layer.

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

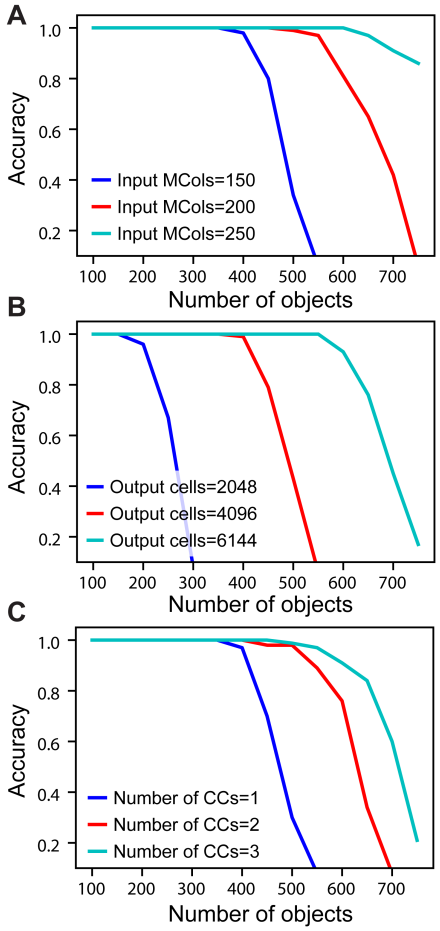

Figure 6. Retrieval accuracy is plotted as a function of the number of learned objects. **A**. Network capacity relative to number of minicolumns in the input layer. The number of output cells is kept at 4096 with 40 cells active at any time. **B**. Network capacity relative to number of cells in the output layer. The number of active output cells is kept at 40. The number of minicolumns in the input layer is 150. C. Network capacity for one, two, and three cortical columns (CCs). The number of minicolumns in the input layer is 150, and the number of output cells is 4096.

MAPPING TO BIOLOGY

Anatomical evidence suggests that the sensorimotor inference model described above exists at least once in each column (layers 4 and 2/3) and perhaps twice (layers 6a and 5). We show the anatomical evidence here and discuss why there might be two circuits in the Discussion section.

Layers 4 and 2/3

The primary instance of the model involves layers 4 and 2/3 as illustrated in **Figure 7A**. The following properties evident in L4 and L2/3 match our model. L4 cells receive direct thalamic input from sensory "core" regions (e.g., LGN) (Douglas and Martin, 2004). This input onto proximal dendrites exhibits driver properties (Viaene et al., 2011a). L4 cells do not form long range connections within their layer (Luhmann et al., 1990). L4 cells project to and activate cells in L2/3 (Lohmann and Rörig, 1994; Feldmeyer et al., 2002; Sarid et al., 2007), and receive feedback from L2/3 (Lefort et al., 2009; Markram et al., 2015). L2/3 cells project long distances within their layer (Stettler et al., 2002; Hunt et al., 2011) and are also a major output of cortical columns (Douglas and Martin, 2004; Shipp et al., 2007).

The model predicts that a representation of location is input to the distal dendrites of the input layer. This is illustrated by the red line in **Figure 7A**. About 45% of L4 synapses come from cells in L6a (Binzegger et al., 2004). The axon terminals were found to show a strong preference for contacting basal dendrites (McGuire et al., 1984) and activation of L6a cells caused weak excitation of L4 cells (Kim et al., 2014). Therefore, we propose that the location representation needed for the upper model comes from L6a.

Layers 6a and 5

Another potential instance of the model is in layers 6a and 5 as illustrated in **Figure 7B**. The following properties evident in L6a and L5 match our model. L6a cells receive direct thalamic input from sensory "core" regions (e.g., LGN) (Thomson, 2010). This input exhibits driver properties and resembles the thalamocortical projections to L4 (Viaene et al., 2011b). L6a cells project to and activate cells in L5 (Thomson, 2010). Recent experimental studies found that the axons of L6 CT neurons densely ramified within layer 5a in both visual and somatosensory cortices of the mouse, and activation of these neurons generated large excitatory postsynaptic potentials (EPSPs) in pyramidal neurons in layer 5a (Kim et al., 2014). L6a cells receive feedback from L5 (Thomson, 2010). L5 cells project long distances within their layer (Schnepel et al., 2015) and L5 cells are also a major output of cortical columns (Douglas and Martin, 2004; Guillery and Sherman, 2011; Sherman and Guillery, 2011). There are three types of pyramidal neurons in L5 (Kim et al., 2015). Here we are referring to only one of them, the larger neurons with thick apical trunks that send an axon branch to relay cells in the thalamus (Ramaswamy and Markram, 2015). However, there is also empirical evidence our model does not map cleanly to L6a and L5. For example, (Constantinople and Bruno, 2013) have shown a sensory stimulus will often cause L5 cells to fire simultaneously or even slightly before L6 cells, which is inconsistent with the model. Therefore, whether L6a and L5 can be interpreted as an instance of the model is unclear.

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

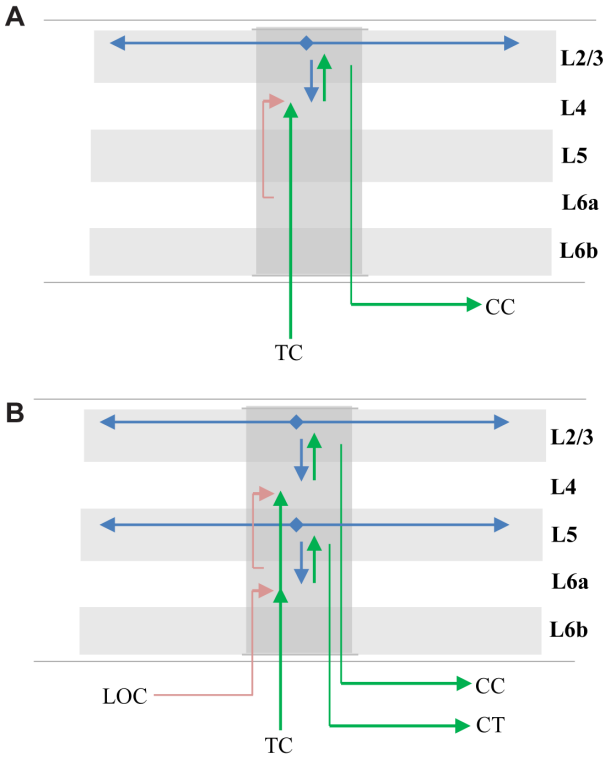

Figure 7. Mapping of sensorimotor inference network onto experimentally observed cortical connections. Arrows represent documented pathways. **A**. First instance of network; L4 is input layer, L2/3 is output layer. Green arrows are feedforward pathway, from thalamo-cortical (TC) relay cells, to L4, to L2/3 cortico-cortical (CC) output cells. Cells in L2/3 also project back to L4 and to adjacent columns (blue arrows); these projections depolarize specific sets of cells that act as predictions (see text). Red arrow is location signal originating in L6a and terminating on basal distal dendrites of L4 cells. **B**. Possible second instance of network; L6a is input layer, L5 is output layer. Both instances of the network receive feedforward input from the same TC axons, thus the two networks run in parallel (Constantinople and Bruno, 2013; Markov et al., 2013). The origin and derivation of the location signal (LOC) is unknown but likely involves local processing as well as input from other regions (see text and Discussion). The output of the upper network makes direct cortical-cortical (CC) connections, whereas the output of the lower network projects to thalamic relay cells before projecting to the next region.

Origin of location signal

The model predicts that a representation of location is input to the distal dendrites of the input layer. This is illustrated by the red lines in **Figure 7**. About 45% of L4 synapses come from cells in L6a (Binzegger et al., 2004) (upper red line in **Figure 7A**). The axon terminals were found to show a strong preference for contacting basal dendrites (McGuire et al., 1984) and activation of L6a cells caused weak excitation of L4 cells (Kim et al., 2014). Therefore, we propose that the location representation needed for the upper model comes from L6a. The derivation of the location representation in L6a is unknown. Part of the answer will involve local processing within the lower layers of the column and part will likely involve long range connections between corresponding regions in "what" and "where" pathways (Thomson, 2010). Parallel "what" and "where" pathways exist in all the major sensory modalities (Ungerleider and Haxby, 1994; Ahveninen et al., 2006). Evidence suggests that regions in "what" pathways form representations that exhibit increasing invariance to translation, rotation or scale and increasing selectivity to sensory features in object centered coordinates (Rust and DiCarlo, 2010). This effect can be interpreted as forming allocentric representations. In contrast, it has been proposed that regions in "where" pathways form representations in egocentric coordinates (Goodale and Milner, 1992). If an egocentric motor behavior is generated in a "where" region, then a copy of the motor command will need to be sent to the corresponding "what" region where it can be converted to a new predicted allocentric location. The conversion is dependent on the current position and orientation of the object relative to the body. It is for this reason we suggest that the origin of the location signal might involve long-range connections between "where" and "what" regions. In the Discussion section we will describe how the location might be generated.

Physiological evidence

In addition to anatomical support, there are several physiological predictions of the model that are supported by empirical observation. L4 and L6a cells exhibit "simple" receptive fields (RFs) while L2/3 and L5 cells exhibit "complex" RFs (Hubel and Wiesel, 1962; Gilbert, 1977). Key properties of complex cells include RFs influenced by a wider area of sensory input and increased temporal stability. L2/3 cells have receptive fields that are twice the size of L4 cells in the primary somatosensory cortex (Chapin, 1986). A distinct group of cells with large and non-oriented receptive fields were found mostly in layer 5 of the visual cortex (Mangini and Pearlman, 1980; Lemmon and Pearlman, 1981). These properties are consistent with, and observed, in the output layer of our model.

The model predicts that cells in a minicolumn in the input layer (L4 and L6a) will have nearly identical RFs when presented with an input than cannot be interpreted as part of a previously learned object. However, in the context of learned objects, the cells in a minicolumn will differentiate. One key differentiation is that individual cells will respond only in specific contexts. This differentiation has been observed in multiple modalities (Vinje and Gallant, 2002; Yen et al., 2006; Martin and Schröder, 2013; Gavornik and Bear, 2014) .

A particularly relevant version of this phenomenon is "border ownership" (Zhou et al., 2000). Cells which have similar classic receptive fields when presented with isolated edge-like features, diverge and fire uniquely when the feature is part of a larger object. Specifically, the cells fire when the feature is at a particular location on a complex object, a behavior predicted and exhibited by our model. To explain border ownership, researchers have proposed a layer of cells that perform "grouping" of inputs. The grouping cells are stable over time (Craft et al., 2007). The output layer of our model performs this function. "Border ownership" is a form of complex object modeling. It has been observed in both primary and secondary sensory regions (Zhou et al., 2000). We predict that similar properties can be observed in primary and secondary sensory regions for more complex and higher-dimensional objects.

Lee, Carvell, et. al show that enhancement of motor cortex activity facilitates sensory-evoked responses of topographically aligned neurons in primary somatosensory cortex (Lee et al., 2008). Specifically, they found that S1

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

corticothalamic neurons in whisker/barrel cortex responded more robustly to whisker deflections when motor cortex activity was focally enhanced. This supports the model hypothesis that behaviorally-generated location information projects in a column-by-column fashion to primary sensory regions.

DISCUSSION

Relationship with previous models

Due to the development of new experimental techniques, knowledge of the laminar circuitry of the cortex continues to grow (Thomson and Bannister, 2003; Thomson and Lamy, 2007). It is now possible to reconstruct and simulate the circuitry in an entire cortical column (Markram et al., 2015). Over the years, numerous efforts have been undertaken to develop models of cortical columns. Many cortical column models aim to explain neurophysiological properties of the cortex. For example, based on their studies on the cat visual cortex, (Douglas and Martin, 1991) provided one of the first canonical microcircuit models of a cortical column. This model explains intracellular responses to pulsed visual stimulations and remained highly influential in the neuroscience community (Douglas and Martin, 2004). (Hill and Tononi, 2004) constructed a large-scale model of point neurons that are organized in a repeating columnar structure to explain the difference of brain states during sleep and wakefulness. (Traub et al., 2004) developed a single-column network model based on multi-compartmental biophysical models to explain oscillatory, epileptic and sleeplike phenomena. (Reimann et al., 2013) showed that the neocortical local field potentials can be explained by a cortical column model composed of >12,000 reconstructed multi-compartmental neurons.

Although these models provided important insights on the origin of neurophysiological signals, there are relatively few models proposing the functional roles of layers and columns. (Bastos et al., 2012) discussed the correspondence between the micro-circuitry of the cortical column and the connectivity implied by predictive coding. This study used a coarse microcircuit model based on the work of (Douglas and Martin, 2004) and lacked recent experimental evidence and detailed connectivity patterns across columns. (Raizada and Grossberg, 2003) described the LAMINART model to explain how attention might be implemented in the visual cortex. This study highlighted the anatomical connections of the L4-L2/3 network and proposed that perceptual grouping relies on long-range lateral connections in L2/3. This is consistent with our proposal of the stable object representation in L2/3.

Generating the location signal

A key prediction of our model is the presence of a location signal in each column of a cortical region. We deduced the need for this signal based on the observation that cortical regions predict new sensory inputs due to movement (Duhamel et al., 1992; Nakamura and Colby, 2002; Li and DiCarlo, 2008). To predict the next sensory input, a patch of neocortex needs to know where a sensor will be on a sensed object after a movement is completed. The prediction of location must be done separately for each part of a sensor array. For example, for the brain to predict what each finger will feel on a given object, it has to predict a separate allocentric location for each finger. There are dozens of semi-independent areas of sensation on each hand, each of which can sense a different part of an object. Thus, allocentric locations must be computed at a fine granular level, and therefore, locations must be computed in a part of the brain where somatic topology is similarly granular. For touch, this suggests location computations are occurring throughout primary regions such as S1 and S2. The same arguments hold for primary visual regions as each patch of the retina observes different parts of objects.

Although we don't know how the location signal is generated, we can list some theoretically-derived requirements. A column needs to know its location on an object, but it also needs to predict what its new location will be after a movement is completed. To translate an egocentric motor signal into a predicted allocentric location, a column must also know the orientation of the object relative to the body part doing the moving: current location + orientation of object + movement => predicted new location. This is a complicated task for neurons to perform. Fortunately, it is highly analogous to what grid cells do. Grid cells are a proof that neurons can perform these types of transformations and they suggest specific mechanisms that might be deployed in cortical columns.

(1) Grid cells in the entorhinal cortex (Hafting et al., 2005; Moser et al., 2008) encode the location of an animal's body relative to an external environment. A sensory column needs to encode the location of a *part* of the animal's body (a sensory patch) relative to an external *object*.

(2) Grid cells use path integration to predict a new location due to movement (Kropff et al., 2015). A column must also use path integration to predict a new location due to movement.

(3) To predict a new location, grid cells combine current location, with movement, with head direction cells (Moser et al., 2014). Head direction cells represent the "orientation" of the "animal" relative to an external environment. Columns need a representation of the "orientation" of a "sensory patch" relative to an external object.

These analogs, plus the fact that grid cells are phylogenetically older than the neocortex, lead us to hypothesize that the cellular mechanisms used by grid cells were preserved and replicated in the sub-granular layers of each cortical column.

Today we have no direct empirical evidence to support this hypothesis. We have only indirect evidence. For example, to compute location, cortical columns must receive dynamically updated inputs regarding body pose. There is now significant evidence that cells in numerous cortical areas, including sensory regions, are modulated by body movement and position. Primary visual and auditory regions contain neurons that are modulated by eye position (Trotter and Celebrini, 1999; Werner-Reiss et al., 2003) as do areas MT, MST, and V4 (Bremmer, 2000; DeSouza et al., 2002). Cells in frontal eye fields (FEF) respond to auditory stimuli in an eye-centered frame of reference (Russo and Bruce, 1994). Posterior parietal cortex (PPC) represents multiple frames of reference including head-centered (Andersen et al., 1993) and body-centered (Duhamel et al., 1992; Brotchie et al., 1995, 2003; Bolognini and Maravita, 2007) representations. Motor areas also contain a diverse range of reference frames, from representations of external space independent of body pose to representations of specific groups of muscles (Graziano and Gross, 1998; Kakei et al., 2003). Many of these representations are granular,

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

specific to particular body areas, and multisensory, implying numerous transformations are occurring in parallel (Graziano et al., 1997; Graziano and Gross, 1998; Rizzolatti et al., 2014). Some models have shown that the above information can be used to perform coordinate transformations (Zipser and Andersen, 1988; Pouget and Snyder, 2000).

Determining how columns derive the allocentric location signal is a current focus of our research. It is also something that should be empirically testable.

Hierarchy

The neocortex processes sensory input in a series of hierarchically arranged regions. As input ascends from region to region, cells respond to larger areas of the sensory array and to more complex features. A common assumption is that complete objects can only be recognized at a level in the hierarchy where cells respond to input over the entire sensory array. For example, it is often said that visual regions V1 and V2 are feature extractors and are incapable of recognizing complete objects, whereas, in higher regions, cells respond to input over most of the retina and therefore object recognition can occur.

Our model proposes an alternate view, that all cortical columns, even columns in primary sensory regions, are capable of learning representations of complete objects. This, however, does not change the basic assumptions regarding hierarchical processing. Our network model is limited by the spatial extent of the horizontal connections in the output layer, and therefore, hierarchy is still required in many situations. For example, say we present an image of a printed letter on the retina, if the letter occupies a small part of the retina, then columns in V1 could recognize the letter. By "recognize the letter" we mean there would be a unique representation of the letter in L2/3 that is stable over eye movements (as long as fixation points remain on or near the letter). If, however, the letter is expanded to occupy a large part of the retina, then columns in V1 would no longer be able to recognize the letter, because the features that define the letter are too far apart to be integrated by the horizontal connections in L2/3. In this case, a converging input onto a subsequent region would be required to recognize the letter. This implies that the cortex learns many models of commonly observed objects. Adjacent columns in a region will learn models of the same objects, as will columns in other regions. Long-range inter-region connections, that start and terminate in the same layer, will allow these models to assist each other in the same way as the long-range intra-laminar connections in the output layer of a single region. We have not yet attempted to incorporate hierarchy in our models, this is an area for future research.

Testable predictions

A number of experimentally testable predictions follow from this theory.

(1) The theory predicts that sensory regions will contain cells that are stable over movements of a sensor while sensing a familiar object.

(2) The set of stable cells will be both sparse and specific to object identity. The cells that are stable for a given object O1 will in general have very low overlap with those that are stable for a completely different object O2.

(3) Activity within the output layer of each column (layers 2/3 and 5) will become sparser as more evidence is accumulated for an object. Activity in the output layer will be denser for ambiguous objects. These effects will only be seen when the animal is freely observing familiar objects.

(4) It is these output layers that form stable representations. In general, their activity will be more stable than layers without long-range connections. Activity within the output layers will converge on a stable representation slower with long-range lateral connections disabled, or by disabling input to adjacent columns.

(5) The theory provides an algorithmic explanation for border ownership cells (Zhou et al., 2000). In general each region will contain cells tuned to the location of features in the object's reference frame. We expect to see these representations in layer 4.

Summary

Our research has focused on how the brain makes predictions of sensory inputs. Starting with the premise that all sensory regions make predictions of their constantly changing input, we deduced that each small area in a sensory region must have access to a location signal that represents where on an object the column is sensing. Building on this idea, we deduced the probable function of several cellular layers and are beginning to understand what cortical columns in their entirety might be doing. Although there are many things we don't understand, the big picture is increasingly clear. We believe each cortical column learns a model of "its" world, of what it can sense. A single column learns the structure of many objects and the behaviors that can be applied to those objects. Through intra-laminar and long-range cortical-cortical connections, columns that are sensing the same object can resolve ambiguity.

In 1978 Vernon Mountcastle reasoned that since the complex anatomy of cortical columns is similar in all of the neocortex, then all areas of the neocortex must be performing a similar function (Mountcastle, 1978). His hypothesis remains controversial partly because we haven't been able to identify what functions a cortical column performs, and partly because it has been hard to imagine what single complex function is applicable to all sensory and cognitive processes.

The model of a cortical column presented in this paper is described in terms of a sensory regions and sensory processing, but the circuitry underlying our model exists in all cortical regions. Thus, if Mountcastle's conjecture is correct, even high-level cognitive functions, such as mathematics, language, and science would be implemented in this framework. It suggests that even abstract knowledge is stored in relation to some form of "location" and that much of what we consider to be "thought" is implemented by inference and behavior generating mechanisms originally evolved to move and infer with fingers and eyes.

MATERIALS AND METHODS

Here we formally describe the activation and learning rules for the HTM sensorimotor inference network. We use a modified version of the HTM neuron model (Hawkins and Ahmad, 2016) in the network. There are three basic aspects of the algorithm: initialization, computing cell states, and updating synapses on dendritic segments. These steps are described below along with

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

implementation and simulation details.

Notation: Let N^{in} represent the number of minicolumns in the input layer, M the number of cells per minicolumn in the input layer, N^{out} the number of cells in the output layer and N^{c} the number of cortical columns. The number of cells in the input layer and output layer is MN^{in} and N^{out} respectively for each cortical column. Each input cell receives both the sensory input and a contextual input that corresponds to the location signal. The location signal is a N^{ext} dimensional sparse vector $\boldsymbol{L}$.

Each cell can be in one of three states: active, predictive, or inactive. We use $M \times N^{in}$ binary matrices $\mathbf{A}^{in}$ and $\mathbf{\Pi}^{in}$ to denote activation state and predictive state of input cells and use N^{out} dimensional binary vectors $\mathbf{A}^{out}$ and $\mathbf{\Pi}^{out}$ to denote activation and predictive state of output cells in a cortical column. The concatenated output of all cortical columns is represented as a $N^{out}N^{column}$ dimensional binary vector $\overline{\mathbf{A}}^{out}$. At any point in time there are only a small number of cells active, so these are generally very sparse.

Each cell maintains a single proximal segment and a set of distal segments. Proximal segments contain feedforward connections to that cell. Distal segments contain contextual input. The contextual input acts as a tiebreaker and biases the cell to win. The contextual input to a cell in the input layer is the external location signal $\boldsymbol{L}$. The contextual input to a cell in the output layer comes from other output cells in the same or different cortical columns.

For each dendritic segment, we maintain a set of "potential" synapses between the dendritic segment and other cells in the network that could potentially form a synapse with it (Chklovskii et al., 2004; Hawkins and Ahmad, 2016). Learning is modeled by the growth of new synapses from a set of potential synapses. A "permanence" value is assigned to each potential synapse and represents the growth of the synapse. Potential synapses are represented by permanence values greater than zero. A permanence value close to zero represents an unconnected synapse that is not fully grown. A permanence value greater than the connection threshold represents a connected synapse. Learning occurs by incrementing or decrementing permanence values.

We denote the synaptic permanences of the dth dendritic segment of the ith input cell in the jth minicolumn as a $N^{ext} \times 1$ vector $\mathbf{D}^{ijd,in}$. Similarly, the permanences of the dth dendritic segment of the ith output cell is the $N^{out}N^{c} \times 1$ dimensional vector $\mathbf{D}^{id,out}$.

Output neurons receive feedforward connections from input neurons within the same cortical column. We denote these connections with a $M \times N^{in} \times N^{out}$

tensor $\mathbf{F}$, where f_{ijk} represents the permanence of the synapse between the ith input cell in the jth minicolumn and the kth output cell.

For $\mathbf{D}$ and $\mathbf{F}$, we will use a dot (e.g. $\dot{\mathbf{D}}$) to denote the binary vector representing the subset of potential synapses on a segment (i.e. permanence value above 0). We use a tilde (e.g. $\widetilde{\mathbf{D}}$) to denote the binary vector representing the subset of connected synapses (i.e. permanence value above connection threshold).

Initialization: Each dendritic segment is initialized to contain a random set of potential synapses. $\mathbf{D}^{ijd,in}$ is initialized to contain a random set of potential synapses chosen from the location input. Segments in $\mathbf{D}^{id,out}$ are initialized to contain a random set of potential synapses to other output cells. These can include cells from the same column. We enforce the constraint that a given segment only contains synapses from a single column. In all cases the permanence values of potential synapses are chosen randomly: initially some are connected (above threshold) and some are unconnected.

Computing cell states:

A cell in the input or output layer is predicted if any of its distal segments have sufficient activity:

$$\pi_{ij}^{in} = \begin{cases} 1 \text{ if } \exists_d \left(\mathbf{L} \cdot \widetilde{\mathbf{D}}^{ijd,in} \geq \theta_b^{in}\right) > 0 \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

$$\pi_{i}^{out} = \begin{cases} 1 \text{ if } \exists_d (\overline{\mathbf{A}}^{out} \cdot \widetilde{\mathbf{D}}^{id,out} \geq \theta_b^{out}) > 0 \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

where θ_b^{in} and θ_b^{out} are the activation thresholds of the basal distal dendrite of an input/output cell respectively.

For the input layer, all the cells in a minicolumn share the same feedforward receptive fields. Following (Hawkins and Ahmad, 2016) we assume that an inhibitory process selects a set of s minicolumns that best match the current feedforward input pattern. We denote this winner set as $\mathbf{W}^{in}$. The set of active input layer cells is calculated as follows:

$$a_{ij}^{in} = \begin{cases} 1 \text{ if } j \in \mathbf{W}^{in} \text{ and } \pi_{ij}^{in} > 0 \\ 1 \text{ if } j \in \mathbf{W}^{in} \text{ and } \sum_i \pi_{ij}^{in} = 0 \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

The first conditional expression represents a cell in a winning minicolumn becoming active if it is predicted. If no cell in a minicolumn is predicted, all cells in that minicolumn become active (second conditional).

To determine activity in the output layer we calculate the feedforward and lateral input to each cell. Cells with sufficient feedforward support from the current time step, and the most lateral support from the previous time step become active. The amount of input overlap to the kth output cell is.:

$$o_k^{out,t} = \sum_{i,j} \mathrm{I}[f_{ijk} \geq \theta_c^{out}] a_{ij}^{in,t} \quad (4)$$

The set of output cells with enough feedforward input is computed as:

$$\mathbf{W}^{out,t} = \{k | o_k^{out,t} \geq \theta_p^{out}\} \quad (5)$$

where θ_p^{out} represents the threshold for the proximal dendrites of output cells. We then sort cells in $\mathbf{W}^{out,t}$ based on the number of active distal segments and select the active ones as:

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$a_i^{out,t} = \begin{cases} 1 \text{ if } i \in \mathbf{W}^{out,t} \text{ and } \rho_i^{out,t-1} \geq \xi_{t-1}^{out} \\ 0 \text{ otherwise} \end{cases} \quad (6)$$

where $\rho_i^{out,t-1} = \sum_d \mathrm{I}[\tilde{\mathbf{A}}^{out} \cdot \tilde{\mathbf{D}}^{id,out} \geq \theta_b^{in}]$ represents the number of active basal distal segments in the previous time step, and the sth highest number of active distal segments is denoted as ξ_t^{out}. $I[]$ is the indicator function, and s is the minimum desired number of active neurons. If the number of cells with lateral support is less than s in a cortical column, $\xi_t^{out,c}$ would be zero and all cells with enough feedforward input will become active.

Learning rule: During learning permanences are modified using a Hebbian-like rule that is specific to segments. In general, if a dendritic segment on a cell becomes active and the cell subsequently becomes active, we reinforce that segment (i.e. increase permanence values of active synapses and decrease permanence values of inactive synapses on that segment).

In the input layer if a cell is depolarized via a basal dendrite and subsequently becomes active, we reinforce the dendritic segment that caused the depolarization. If no cell in an active minicolumn is predicted, we select the cell with the most activated segment and perform learning on that segment. When there is no cell with a sufficiently active segment, we choose a random segment on a random cell in that minicolumn to learn. Learning on a segment involves decreasing the permanence of inactive synapses by a small value p^- and increasing the permanence of active synapses by a larger value p^+:

$$\Delta\mathbf{D}^{ijd,in} = p_{in}^{+}\dot{\mathbf{D}}^{ijd,in} \circ \mathbf{L}^{t-1} - p_{in}^{-}\dot{\mathbf{D}}^{ijd,in} \circ (\mathbf{1} - \mathbf{L}^{t-1}) \quad (7)$$

where $\circ$ represents element-wise multiplication.

When learning a new object in the output layer, we randomly select a sparse set of output cells to represent it, and keep those cells active while learning that object. The same learning rule as above is applied to the basal distal dendrites of active output cells.

$$\Delta\mathbf{D}^{id,out} = p_{out}^{+}\dot{\mathbf{D}}^{id,out} \circ \bar{\mathbf{A}}^{out,t-1} - p_{out}^{-}\dot{\mathbf{D}}^{id,out} \circ (\mathbf{1} - \bar{\mathbf{A}}^{out,t-1}) \quad (8)$$

The feedforward connections on proximal segments are learned using a similar Hebbian rule. For each active output cell, we reinforce active input connections by increasing the synaptic permanence by p_{ff}^{+}, and punish inactive connections by decreasing the synaptic permanence by p_{ff}^{-}.

$$\Delta f_{ijk} = [p_{ff}^{+} a_{ij}^{in} - p_{ff}^{-}(1 - a_{ij}^{in})]\mathrm{I}[f_{ijk} > 0] \quad (9)$$

Simulation details: To generate our convergence and capacity results we generated a large number of objects. Each object consists of a number of sensory features, with each feature assigned to a corresponding location. We encode each location as a 2400-dimensional sparse binary vector with 10 random bits active. Each sensory feature is similarly encoded by a vector with 10 random bits active. The length of the sensory feature vector is the same as the number of minicolumns of the input layer N^{in}. The input layer contains 150 mini-columns and 16 cells per mini-column, with 10 mini-columns active at any time. The activation threshold of basal distal dendrite of input neuron is 6. The output layer contains 4096 cells and the minimum number of active output cells is 40. The activation threshold is 3 for proximal dendrites and 18 for distal dendrites for output neurons.

During training, the network learns each object in random order. For each object, the network senses each feature three times. The activation pattern in the output layer is saved for each object to calculate retrieval accuracy. During testing, we allow the network to sense each object at K locations. After each sensation, we classify the activity pattern in the output layer. We say that an object is correctly classified if, for each cortical column, the overlap between the output layer and the stored representation for the correct object is above a threshold, and the overlaps with the stored representation for all other objects are below that threshold. We use a threshold of 30.

For the network convergence experiment (**Figure 5-6**), each object consists of 10 sensory features chosen from a library of 5 to 30 possible features. The number of sensations during testing is 20. For the capacity experiment, each object consists of 10 sensory features chosen from a large library of 5000 possible features. The number of sensations during testing is 3.

Finally, we make some simplifying assumptions that greatly speed up simulation time for larger networks. Instead of explicitly initializing a complete set of synapses across every segment and every cell, we greedily create segments on a random cell and initialize potential synapses on that segment by sampling from currently active cells. This happens only when there is no match to any existing segment.

ACKNOWLEDGEMENTS

We thank Jeff Gavornik for his thoughtful comments and suggestions. We also thank numerous collaborators at Numenta over the years for many discussions.

REFERENCES

Ahmad S, Hawkins J (2016) How do neurons operate on sparse distributed representations? A mathematical theory of sparsity, neurons and active dendrites. arXiv:1601.00720 [q-NC].

Ahveninen J, Jääskeläinen IP, Raij T, Bonmassar G, Devore S, Hämäläinen M, Levänen S, Lin F-H, Sams M, Shinn-Cunningham BG, Witzel T, Belliveau JW (2006) Task-modulated "what" and "where" pathways in human auditory cortex. Proc Natl Acad Sci U S A 103:14608–14613.

Andersen RA, Snyder LH, Li CS, Stricanne B (1993) Coordinate transformations in the representation of spatial information. Curr Opin Neurobiol 3:171–176.

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ (2012) Canonical microcircuits for predictive coding. Neuron 76:695–711.

Binzegger T, Douglas RJ, Martin KAC (2004) A Quantitative Map of the Circuit of Cat Primary Visual Cortex. J Neurosci 24:8441–8453.

Bolognini N, Maravita A (2007) Proprioceptive Alignment of

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Visual and Somatosensory Maps in the Posterior Parietal Cortex. Curr Biol 17:1890–1895.

Bremmer F (2000) Eye position effects in macaque area V4. Neuroreport 11:1277–1283.

Brotchie PR, Andersen RA, Snyder LH, Goodman SJ (1995) Head position signals used by parietal neurons to encode locations of visual stimuli. Nature 375:232–235.

Brotchie PR, Lee MB, Chen DY, Lourensz M, Jackson G, Bradley WG (2003) Head position modulates activity in the human parietal eye fields. Neuroimage 18:178–184.

Buxhoeveden DP (2002) The minicolumn hypothesis in neuroscience. Brain 125:935–951.

Chapin JK (1986) Laminar differences in sizes, shapes, and response profiles of cutaneous receptive fields in the rat SI cortex. Exp Brain Res 62:549–559.

Chklovskii DB, Mel BW, Svoboda K (2004) Cortical rewiring and information storage. Nature 431:782–788.

Constantinople CM, Bruno RM (2013) Deep cortical layers are activated directly by thalamus. Science 340:1591–1594.

Craft E, Schutze H, Niebur E, von der Heydt R (2007) A Neural Model of Figure-Ground Organization. J Neurophysiol 97:4310–4326.

DeSouza JF, Dukelow SP, Vilis T (2002) Eye position signals modulate early dorsal and ventral visual areas. Cereb Cortex 12:991–997.

Douglas RJ, Martin KA (1991) A functional microcircuit for cat visual cortex. J Physiol 440:735–769.

Douglas RJ, Martin KACC (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27:419–451.

Duhamel J, Colby CL, Goldberg ME (1992) The Updating of the Representation of Visual representation. Science (80-) 255:90–92.

Feldmeyer D, Lübke J, Silver RA, Sakmann B (2002) Synaptic connections between layer 4 spiny neurone-layer 2/3 pyramidal cell pairs in juvenile rat barrel cortex: physiology and anatomy of interlaminar signalling within a cortical column. J Physiol 538:803–822.

Gavornik JP, Bear MF (2014) Learned spatiotemporal sequence recognition and prediction in primary visual cortex. Nat Neurosci 17:732–737.

Gilbert CD (1977) Laminar differences in receptive field properties of cells in cat primary visual cortex. J Physiol 268:391–421.

Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends Neurosci 15:20–25.

Graziano MS, Gross CG (1998) Spatial maps for the control of movement. Curr Opin Neurobiol 8:195–201.

Graziano MS, Hu XT, Gross CG (1997) Visuospatial properties of ventral premotor cortex. J Neurophysiol 77:2268–2292.

Guillery RW, Sherman SM (2011) Branched thalamic afferents: what are the messages that they relay to the cortex? Brain Res Rev 66:205–219.

Hafting T, Fyhn M, Molden S, Moser M-B, Moser EI (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436:801–806.

Hawkins J, Ahmad S (2016) Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex. Front Neural Circuits 10.

Hill S, Tononi G (2004) Modeling Sleep and Wakefulness in the Thalamocortical System. J Neurophysiol 93:1671–1698.

Horton JC, Adams DL (2005) The cortical column: a structure without a function. Philos Trans R Soc Lond B Biol Sci 360:837–862.

Hubel D, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154.

Hunt JJ, Bosking WH, Goodhill GJ (2011) Statistical structure of lateral connections in the primary visual cortex. Neural Syst Circuits 1:3.

Kakei S, Hoffman DS, Strick PL (2003) Sensorimotor transformations in cortical motor areas. Neurosci Res 46:1–10.

Kim EJ, Juavinett AL, Kyubwa EM, Jacobs MW, Callaway EM (2015) Three Types of Cortical Layer 5 Neurons That Differ in Brain-wide Connectivity and Function. Neuron 88:1253–1267.

Kim J, Matney CJ, Blankenship A, Hestrin S, Brown SP (2014) Layer 6 corticothalamic neurons activate a cortical output layer, layer 5a. J Neurosci 34:9656–9664.

Kropff E, Carmichael JE, Moser M-B, Moser EI (2015) Speed cells in the medial entorhinal cortex. Nature 523:419–424.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444.

Lee S, Carvell GE, Simons DJ (2008) Motor modulation of afferent somatosensory circuits. Nat Neurosci 11:1430–1438.

Lefort S, Tomm C, Floyd Sarria J-C, Petersen CCH (2009) The Excitatory Neuronal Network of the C2 Barrel Column in Mouse Primary Somatosensory Cortex. Neuron 61:301–316.

Lemmon V, Pearlman AL (1981) Does laminar position determine the receptive field properties of cortical neurons? A study of corticotectal cells in area 17 of the normal mouse and the reeler mutant. J Neurosci 1:83–93.

Li N, DiCarlo JJ (2008) Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science (80-) 321:1502–1507.

Lohmann H, Rörig B (1994) Long-range horizontal connections between supragranular pyramidal cells in the extrastriate visual cortex of the rat. J Comp Neurol 344:543–558.

Losonczy A, Makara JK, Magee JC (2008) Compartmentalized dendritic plasticity and input feature storage in neurons. Nature 452:436–441.

Luhmann HJ, Singer W, Martínez-Millán L (1990) Horizontal Interactions in Cat Striate Cortex: I. Anatomical

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Substrate and Postnatal Development. Eur J Neurosci 2:344–357.

Maass W (1997) Networks of spiking neurons: the third generation of neural network models. Neural networks 10:1659–1671.

Mangini NJ, Pearlman AL (1980) Laminar distribution of receptive field properties in the primary visual cortex of the mouse. J Comp Neurol 193:203–222.

Markov NT, Ercsey-Ravasz M, Van Essen DC, Knoblauch K, Toroczkai Z, Kennedy H (2013) Cortical high-density counterstream architectures. Science 342:1238406.

Markram H et al. (2015) Reconstruction and Simulation of Neocortical Microcircuitry. Cell 163:456–492.

Martin KAC, Schröder S (2013) Functional heterogeneity in neighboring neurons of cat primary visual cortex in response to both artificial and natural stimuli. J Neurosci 33:7325–7344.

McGuire BA, Hornung JP, Gilbert CD, Wiesel TN (1984) Patterns of synaptic input to layer 4 of cat striate cortex. J Neurosci 4:3021–3033.

Moser EI, Kropff E, Moser M-B (2008) Place cells, grid cells, and the brain's spatial representation system. Annu Rev Neurosci 31:69–89.

Moser EI, Roudi Y, Witter MP, Kentros C, Bonhoeffer T, Moser M-B (2014) Grid cells and cortical representation. Nat Rev Neurosci 15:466–481.

Mountcastle V (1978) An organizing principle for cerebral function: the unit model and the distributed system. In: The Mindful Brain (Edelman G, Mountcastle V, eds). Cambridge, Mass.: MIT Press.

Mountcastle VB (1997) The columnar organization of the neocortex. Brain 120:701–722.

Nakamura K, Colby CL (2002) Updating of the visual representation in monkey striate and extrastriate cortex during saccades. Proc Natl Acad Sci U S A 99:4026–4031.

Pouget A, Snyder LH (2000) Computational approaches to sensorimotor transformations. Nat Neurosci 3 Suppl:1192–1198.

Raizada RDS, Grossberg S (2003) Towards a Theory of the Laminar Architecture of Cerebral Cortex: Computational Clues from the Visual System. Cereb Cortex 13:100–113.

Ramaswamy S, Markram H (2015) Anatomy and physiology of the thick-tufted layer 5 pyramidal neuron. Front Cell Neurosci 9:233.

Reimann MW, Anastassiou CA, Perin R, Hill SL, Markram H, Koch C (2013) A Biophysically Detailed Model of Neocortical Local Field Potentials Predicts the Critical Role of Active Membrane Currents. Neuron 79:375–390.

Rizzolatti G, Cattaneo L, Fabbri-Destro M, Rozzi S (2014) Cortical mechanisms underlying the organization of goal-directed actions and mirror neuron-based action understanding. Physiol Rev 94:655–706.

Russo GS, Bruce CJ (1994) Frontal eye field activity preceding aurally guided saccades. J Neurophysiol 71:1250–1253.

Rust NC, DiCarlo JJ (2010) Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. J Neurosci 30:12978–12995.

Sarid L, Bruno R, Sakmann B, Segev I, Feldmeyer D (2007) Modeling a layer 4-to-layer 2/3 module of a single column in rat neocortex: interweaving in vitro and in vivo experimental observations. Proc Natl Acad Sci U S A 104:16353–16358.

Schnepel P, Kumar A, Zohar M, Aertsen A, Boucsein C (2015) Physiology and impact of horizontal connections in rat neocortex. Cereb Cortex 25:3818–3835.

Sherman SM, Guillery RW (2011) Distinct functions for direct and transthalamic corticocortical connections. J Neurophysiol 106:1068–1077.

Shipp S et al. (2007) Structure and function of the cerebral cortex. Curr Biol 17:R443-9.

Spruston N (2008) Pyramidal neurons: dendritic structure and synaptic integration. Nat Rev Neurosci 9:206–221.

Stettler DD, Das A, Bennett J, Gilbert CD (2002) Lateral connectivity and contextual interactions in macaque primary visual cortex. Neuron 36:739–750.

Stuart GJ, Häusser M (2001) Dendritic coincidence detection of EPSPs and action potentials. Nat Neurosci 4:63–71.

Thomson AM (2010) Neocortical layer 6, a review. Front Neuroanat 4:13.

Thomson AM, Bannister AP (2003) Interlaminar connections in the neocortex. Cereb Cortex 13:5–14.

Thomson AM, Lamy C (2007) Functional maps of neocortical local circuitry. Front Neurosci 1:19–42.

Traub RD, Contreras D, Cunningham MO, Murray H, LeBeau FEN, Roopun A, Bibbig A, Wilent WB, Higley MJ, Whittington MA (2004) Single-Column Thalamocortical Network Model Exhibiting Gamma Oscillations, Sleep Spindles, and Epileptogenic Bursts. J Neurophysiol 93:2194–2232.

Trotter Y, Celebrini S (1999) Gaze direction controls response gain in primary visual-cortex neurons. Nature 398:239–242.

Ungerleider LG, Haxby J V (1994) "What" and "where" in the human brain. Curr Opin Neurobiol 4:157–165.

Viaene AN, Petrof I, Sherman SM (2011a) Synaptic properties of thalamic input to layers 2/3 and 4 of primary somatosensory and auditory cortices. J Neurophysiol 105:279–292.

Viaene AN, Petrof I, Sherman SM (2011b) Synaptic properties of thalamic input to the subgranular layers of primary somatosensory and auditory cortices in the mouse. J Neurosci 31:12738–12747.

Vinje WE, Gallant JL (2002) Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1. J Neurosci 22:2904–2915.

Werner-Reiss U, Kelly KA, Trause AS, Underhill AM, Groh JM (2003) Eye position affects activity in primary auditory cortex of primates. Curr Biol 13:554–562.

bioRxiv preprint doi: https://doi.org/10.1101/162263; this version posted July 12, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Yen S-C, Baker J, Gray CM (2006) Heterogeneity in the Responses of Adjacent Neurons to Natural Stimuli in Cat Striate Cortex. J Neurophysiol 97:1326–1341.

Zhou H, Friedman HS, von der Heydt R (2000) Coding of border ownership in monkey visual cortex. J Neurosci 20:6594–6611.

Zipser D, Andersen RA (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331:679–684.