Deep Neocortical Dynamics for Predictive Learning and Attention

Randall C. O'Reilly, et al
Department of Psychology and Neuroscience
University of Colorado Boulder
345 UCB
Boulder, CO 80309
randy.oreilly@colorado.edu

January 30, 2016

Submitted Manuscript: Do not cite or quote without permission.
Supported by: ONR grant N00014-13-1-0067, ONR N00014-10-1-0177, ONR D00014-12-C-0638, and Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

Abstract

todo: old: We present a comprehensive, novel framework for understanding how the neocortex, including the thalamocortical loops through the deep layers, can support a temporal context representation in the service of predictive learning. Many have argued that predictive learning provides a compelling, powerful source of learning signals to drive the development of human intelligence: if we constantly predict what will happen next, and learn based on the discrepancies from our predictions (error-driven learning), then we can learn to improve our predictions by developing internal representations that capture the regularities of the environment (e.g., physical laws governing the time-evolution of object motions). Our version of this idea builds upon existing work with simple recurrent networks (SRN's), which have a discretely-updated temporal context representations that are a direct copy of the prior internal state representation. We argue that this discretization of temporal context updating has a number of important computational and functional advantages, and further show how the strong alpha-frequency (10hz, 100ms cycle time) oscillations in the posterior neocortex could reflect this temporal context updating. Specifically, layer 5b intrinsically bursting neurons fire at the alpha frequency, and trigger an updating of the layer 6 regular spiking neurons that project down to the thalamus, and from there go back up to layer 4 and layer 6 – this thalamocortical loop sustains the temporal context representation as the system develops a prediction about what will happen next. When *next* inevitably happens, any resulting discrepancy can drive a biologically-based form of error-driven learning, which we have developed as part of the Leabra framework. Thus, we refer to this new model as LeabraTI (temporal integration). We examine a wide range of data from biology to behavior through the lens of this LeabraTI model, and find that it provides a unified account of a number of otherwise disconnected findings, all of which converge to support this new model of neocortical learning and processing. We describe an implemented model showing how predictive learning of tumbling object trajectories can facilitate object recognition with cluttered backgrounds.

Introduction

How does the neocortex support the remarkable learning mechanisms that result, after many years of experience, in our considerable perceptual and cognitive abilities? Answering this central question has been the holy grail of many lines of research, at many levels of analysis from synapses up to machine learning algorithms. Despite many advances at each of these levels, we still lack a suitable framework for outlining the key elements of a comprehensive answer to this question that has the potential to integrate across these different levels in a mutually compatible way. In this paper, we attempt to articulate such a framework, which we think provides a broad and deep integration of many different sources of data and theoretical ideas coming from many different researchers, and is implemented in computer models that demonstrate both its computational power and its ability to account for a wide range of data. Although this framework is focused on understanding the nature of learning in the neocortex, learning can only be understood in the context of the information processing machine in which it operates, and capturing the central principles and mechanisms of neocortical information processing represents a tremendous challenge for any such attempt. Clearly, given the incredible complexity and lurking uncertainties in our present state of knowledge, any such framework must be considered entirely provisional at this point, but we feel that the present state of our efforts represents a sufficient increment of progress that the time is ripe for a major explication of the current state of progress, along with a number of important future directions that will need to be addressed in future work.

To make our objectives and central principles concrete, we focus on a task that has been widely recognized as tapping the central elements of general intelligence (represented by the g factor in psychometric studies of intelligence), the *Raven's* progressive matricies task (*RPM*; Figure 1; Raven, 1941, 2000). First, this task has a strong perceptual element to it, requiring complex visual displays composed of many different elements to be parsed appropriately. Thus, a suitable model of learning in this domain must explain how the human visual system learns to transform the basic oriented edge detectors and other such features in V1 into higher-level perceptual representations that can support performance on such a task. But that is just the starting point for the more challenging problem of understanding how people learn to recognize sequential patterns across the different elements in the rows and columns of such displays, and how they can flexibly formulate different hypotheses about which patterns seem sufficient to account for the existing stimuli, and what this then implies for the key missing element in the lower right-hand corner. How are the plans and strategies learned that govern behavior at this level, and how do these interact with the lower-level perceptual representations such that we can focus our attention on different stimulus dimensions according to the higher-level hypotheses we're exploring? Clearly, a computational model that truly answered such questions at a deep and satisfying level, providing a naturalistic explanation for how such core perceptual and cognitive abilities arise from the everyday experiences of people in realistic environments, would be a tremendous accomplishment, heralding the advent of true artificial intelligence approaching the core capabilities of the human brain. We certainly have not yet achieved this goal, but we do think that the framework articulated and demonstrated here has, for the first time, a clear pathway toward achieving such a goal.

To see why we think this, we unpack the above summary of the Raven's task and examine each of the elements in greater detail, and in relationship to existing computational models compared to our new framework.

In the domain of perceptual learning, there have been considerable advances in object recognition performance at the machine learning level, that nicely build upon and integrate central ideas and data from perceptual neuroscience. The current *convolutional neural network* (CNN) models trained by the error backpropagation learning mechanism (Ciresan, Meier, Gambardella, & Schmidhuber, 2010; Ciresan, Meier, & Schmidhuber, 2012; Bengio, Courville, & Vincent, 2013) develop deep hierarchies of increasingly complex and spatially invariant feature detectors, that ultimately support high-accuracy recognition of large numbers

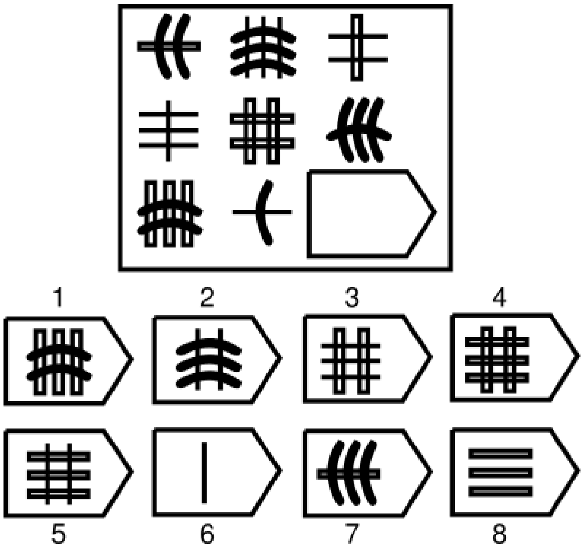

Figure 1: The Raven's progressive matricies task. The objective is to find the appropriate element for the lower-right-hand cell, which completes the patterns established by the existing rows, columns, and diagonals. This task requires extensive flexible inspection of the separable visual features of the elements, and integration of features across elements to develop patterns. Computationally, this requires dynamic top-down and bottom-up interactions in visual and higher-level cognitive processing, with flexible attentional highlighting of relevant features, all operating on learned perceptual representations that are sufficiently flexible to allow these features to be extracted and combined in multiple ways, under strategic control. Pretty sure the answer is 5.

of objects in real-world visual images. The qualitative (and to some extent quantitative) nature of these deep hierarchies maps reasonably well onto what we know about the visual pathways going from V1 up to inferotemporal (IT) cortex (e.g., Majaj, Hong, Solomon, & DiCarlo, 2015) — when combined with earlier neuroscience-inspired models of these pathways (e.g., Riesenhuber & Poggio, 1999), a fairly compelling understanding across computational and neuroscience levels seems to be emerging.

However, most of these existing models are strictly feedforward in their dynamics, and are thus incapable of accounting for the dynamic, bidirectional interplay of bottom-up perception and top-down attention that is essential for the Raven's task, and for most real-world cognitive tasks that build upon perceptual information. We firmly believe that a core feature of human perception and perceptual learning is that it is driven fundamentally by top-down, goal-driven forces, and that perception is an intrinsically dynamic, active process that is not well captured by simple feedforward models. Thus, our framework shares many core principles with a number of theorists who similarly emphasize active, dynamic visual processing lying at the core of an embodied, sensory-motor foundation to higher-level cognitive function (Barsalou, 2008, 2009; Anderson, 2003). Furthermore, our framework is strongly consistent with Lamme's ideas regarding the nature and importance of recurrent, bidirectional dynamics and consciousness (Lamme, 2006) – a major function of consciousness under this view is to support flexible top-down access and manipulation of bottom-up perceptual information.

Computationally, our framework builds upon existing work developing models of the bidirectional dynamics in deep object recognition hierarchies, where the essential error-driven learning mechanism depends on bidirectional activation dynamics as well (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013; Wyatte, Herd, Mingus, & O'Reilly, 2012b; Wyatte, Curran, & O'Reilly, 2012a). This work is based on the *Leabra* framework (local, error-driven and associative, biologically realistic algorithm), which is pronounced like the *libra* balance scales and represents the attempt to strike a middle-ground balance between biological and computational considerations, along with a healthy balance of multiple forms of learning (OReilly, Hazy, & Herd, 2015; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; O'Reilly & Munakata, 2000; O'Reilly, 1996). Our new framework goes much further in capturing the biological and functional properties of attentional dynamics by incorporating the abstract computational framework of Reynolds and Heeger (2009) while leveraging the insights of Grossberg (1999) and others (Montijn, Klink, & Van Wezel, 2012) regarding the central role of the deep neocortical layers (layers 5-6) and thalamocortical loops in supporting these attentional dynamics. Thus, we call this new framework *DeepLeabra*, because it integrates the functions of the deep neocortical layers, along with deep hierarchical structure across layers, into the core mechanisms associated with the original Leabra framework, which captures some of the core functions of the superficial cortical layers.

In summary, our new framework provides a rich set of activation dynamics within which the learning mechanisms operate, supporting fluid, dynamic allocation of attention in response to both bottom-up and top-down factors, with the superficial and deep neocortical layers each playing distinct but interrelated functional roles. Specifically, the superficial layers engage in high-bandwidth, higher-frequency continuous bidirectional interactions, that can be computationally characterized in terms of *parallel constraint satisfaction* and *attractor dynamics* in the classic tradition of the Hopfield network (Hopfield, 1982, 1984), Boltzmann machine (Ackley, Hinton, & Sejnowski, 1985), and interactive-activation and competition (IAC) (Rumelhart & McClelland, 1982) models. This characterization is consistent with recent evidence showing that superficial cortical layers exhibit temporal synchrony dynamics in the gamma frequency band (around 40Hz), in contrast to the deep cortical layers that exhibit lower, alpha frequency (around 10Hz) temporal synchrony dynamics (Buffalo, Fries, Landman, Buschman, & Desimone, 2011). Correspondingly, the deep layers in our framework function as a kind of outer-loop to the superficial-layer inner loop of processing, imposing more slowly-evolving attentional constraints on the superficial-layer dynamics. The deep layers then update roughly every 100 msec to reflect the evolving "understanding" emerging in the superficial layer representations, along with top-down task/goal constraints coming from higher layers. Computationally, the entire superficial-deep dynamic can be understood in terms of the expectation-maximization (EM) algorithm, which has the same nested inner-outer loop structure (Dempster, Laird, & Rubin, 1977).

How does learning then shape these superficial / deep activation dynamics, in a way that leads to powerful, task-relevant perceptual and higher-level conceptual representations? We have long argued that error-driven learning is the only form of learning that is sufficiently powerful to shape such representations (O'Reilly, 1996, 1998; O'Reilly & Munakata, 2000; O'Reilly et al., 2012; OReilly et al., 2015) and the recent successes of the deep object recognition models cited above would seem to reinforce this notion. However, there has been a long history of skepticism that error-driven learning is biologically realistic (e.g., Crick, 1989), and a focus on Hebbian learning mechanisms in the domain of synaptic plasticity and biologically-realistic neural models. We have attempted to bridge this gap by showing that spike-timing dependent plasticity (STDP) (Bi & Poo, 1998; Markram, Lubke, & Sakmann, 1997) Hebbian learning mechanisms, when modeled in detailed mechanistic fashion (Urakubo, Honda, Froemke, & Kuroda, 2008), yield a learning rule that can support both error-driven and Hebbian learning (O'Reilly et al., 2012; OReilly et al., 2015). This learning rule is in the form of the *BCM* algorithm (Bienenstock, Cooper, & Munro, 1982; Cooper, Intrator, Blais, & Shouval, 2004; Shouval, Wang, & Wittenberg, 2010), where the floating threshold in the BCM algorithm can adapt on a fast-enough time scale to represent the *minus phase* of the contrastive-hebbian error-driven learning rule (CHL), which in turn can be derived directly from error-backpropagation (O'Reilly, 1996) and from the Boltzmann machine learning rule (Ackley et al., 1985; Hinton, 1989). Recent in-vivo tracking of synaptic plasticity supports exactly this form of learning rule, with a BCM-like shape and evidence of rapid movement of the floating threshold (Lim, McKee, Woloszyn, Amit, Freedman, Sheinberg, & Brunel, 2015).

The above arguments provide some powerful raw ingredients for a biologically and cognitively realistic learning mechanism, but major questions remain. The most important of these (to us) is the question of where the target or *plus phase* training signals come from to support error-driven learning. In the current deep object recognition models (and our Leabra versions thereof), the target signals are the category labels of the objects in the images. It is plausible that some reasonable sample of visual experiences are accompanied by auditory signals of object names, but certainly this is a small fraction of the total visual experiences a child has (Yu & Smith, 2012; Smith, Suanda, & Yu, 2014). Thus, our new framework incorporates the widely-explored idea that people learn continuously from each moment of experience through *predictive learning*: learning from the differences between *expectations* versus actual *outcomes* coming in through the sensory input at every moment (Elman, 1990, 1991; Jordan, 1989; Schuster & Paliwal, 1997; Hawkins & Blakeslee, 2004; George & Hawkins, 2009). This form of learning does not require special coincidences of inputs, and instead operates continuously on bottom-up sensory input to develop increasingly accurate internal models of the environment. Our specific proposal is that predictive learning operates within the 10Hz alpha-frequency dynamics of the deep layers and thalamocortical loops (Lorincz, Kekesi, Juhasz, Crunelli, & Hughes, 2009; Franceschetti, Guatteo, Panzica, Sancini, Wanke, & Avanzini, 1995; Buffalo et al., 2011; Luczak, Bartho, & Harris, 2013), with each 100 msec iteration of deep layer / thalamic updating supporting one expectation / outcome predictive learning cycle. The idea that the brain imposes an internal discretization of experience for the purposes of learning (and attentional updating) helps to resolve the otherwise thorny problem of how the brain knows the difference between the minus and plus phases of CHL-style error-driven learning: it is just built-in to the basic dynamics. The resulting overall learning rule incorporates an auto-encoder dynamic in addition to the predictive learning dynamic, thus capturing another widely-embraced and explored computational learning principle known variously as *recognition by synthesis* or *generative model* based learning (e.g., Rumelhart & McClelland, 1986; Carpenter & Grossberg, 1987; Pollack, 1990; Dayan, Hinton, Neal, & Zemel, 1995; Rao & Ballard, 1999; Hinton & Salakhutdinov, 2006; Friston, 2005, 2010; Bengio et al., 2013).

In summary, the DeepLeabra framework integrates an extensive swath of biological data on the superficial and deep / thalamic networks of the neocortex, and considerable data on the biology of synaptic plasticity, to support high-performance error-driven and Hebbian learning based on the principles of predictive auto-encoder learning, within complex bidirectional network dynamics that support top-down and bottom-up attentional and constraint satisfaction processing. We review a broad range of human neuroimaging and behavioral data that fits remarkably well within this framework, including the striking evidence showing that perception is discretized (at least to some extent) at the alpha frequency (VanRullen & Koch, 2003). Overall, we find that this learning framework integrates a wide range of previously unconnected biological and behavioral data, under a coherent, computationally powerful model. The detailed properties of the specialized neuron types within each of the different neocortical layers, and the thalamus, fit well with the multiple different computational functions required by this framework, and we find that we can embed multiple different such functions synergistically within the same neural hardware, making for a very efficient and elegant account of the elaborate structure of the neocortex.

In addition to reviewing and synthesizing this diverse body of empirical data, we show how our model can account for a range of specific data on attentional dynamics in visual cortex, and then demonstrate an initial attempt to simulate core elements of the Raven's progressive matricies task using ecologically-realistic learned representations and appropriate attentional dynamics evolving over time, and we demonstrate the computational power of this model in the context of object recognition with cluttered visual displays. To limit the scope of this paper and the complexity of our models, we exclude detailed consideration of the contributions of the prefrontal cortex, basal ganglia, and hippocampus to shaping the learning and dynamics of human cognition – we have addressed these issues in other papers, and will integrate them within this new framework in future work.

In the remainder of this paper, we introduce the specific ideas for how the thalamocortical and neocortical laminar structure supports temporal integration and learning, and then review the relevant empirical literature across the biology and behavioral domains that bears on the specific computationally and biologically-motivated claims of the DeepLeabra model. Then, we present an application of the model to object recognition in cluttered visual scenes, and conclude with a discussion including comparisons with other related approaches.

Superficial and Deep Neocortical Layer Dynamics for Attention and Learning

We begin with a summary overview of how the DeepLeabra model works, in terms of differential functional roles for superficial and deep layers of the neocortex, and loops through the thalamus, and the temporal dynamics of information flow through this circuit. Then, we explore each of these elements in greater depth, in relation to available biological and cognitive data.

Overview of DeepLeabra Model

The central hypotheses in our framework in broad, overview form are:

- The neocortex is composed of two separable but tightly interacting sub-networks, superficial and deep / thalamic. The superficial-layer network consists of neocortical layers 4, 2, and 3, across different brain areas, with extensive bidirectional interconnectivity (feedforward going from 2/3 to layer 4 in the next layer, and feedback coming from 2/3 in one area back to 2/3 in an earlier area; Markov, Vezoli, Chameau, Falchier, Quilodran, Huissoud, Lamy, Misery, Giroud, Ullman, Barone, Dehay, Knoblauch, & Kennedy, 2014). The deep / thalamic network starts in each area with the layer 5b intrinsic bursting (IB) neurons (5IB, Connors, Gutnick, & Prince, 1982; Sherman & Guillery, 2006), which receive inputs from local superficial neurons and top-down projections from other areas (e.g., higher-level task control signals). These 5IB neurons then project to deep layer 6, which interconnects with the thalamus (which in turn projects back up to layer 4 of the superficial network and layer 6 in the deep network), and the 5IB neurons also provide a strong driving feedforward input to higher-layer thalamic areas.

- The superficial network can be described computationally in terms of a classic Hopfield network / Boltzmann machine constraint satisfaction system (Hopfield, 1982, 1984; Ackley et al., 1985; Rumelhart & McClelland, 1982), that settles over many bidirectional activation propagation updates into a state (representation) that best satisfies the current bottom-up inputs and top-down knowledge / task-driven constraints. This does not imply that the network converges fully to a stable settled attractor state – just that it moves in that direction within the relevant time frame (100 msec as described next), after which changes in the deep / thalamic network (and in the sensory inputs) drive a new settling process under new constraints.

- The deep / thalamic network in the posterior cortex updates at the alpha frequency (roughly every 100 msec), which is the intrinsic bursting frequency of the layer 5IB neurons, and thalamocortical networks also entrain at this frequency due to various mechanisms (Lorincz et al., 2009; Franceschetti et al., 1995; Buffalo et al., 2011; Luczak et al., 2013). Thus, the deep / thalamic network provides a relatively stable set of inputs to the superficial network over this 100 msec time period – the deep state is sustained through regular spiking layer 6 neurons (i.e., layer 6CT corticothalamic neurons; Thomson, 2010; Thomson & Lamy, 2007) that project to the thalamic relay cells (TRC) of the thalamus, which project back to these same 6CT neurons, as well as up to the layer 4 inputs to the superficial network.

- Computationally, the deep / thalamic network activations encode both attentional modulations of the superficial layer state, and temporal context information that reflects activations from the prior 100 msec period. The attentional modulation signals cause the iterative constraint satisfaction process in the superficial network to focus on task-relevant information while down-regulating responses to irrelevant information, consistent with the abstract Reynolds and Heeger (2009) model, while the contributions of the deep layer networks to this function are broadly consistent with the folded-feedback model (Grossberg, 1999). Biologically, the layer 6CT neurons are known to exhibit a multiplicative influence over firing of superficial-layer neurons, in a manner consistent with the Reynolds and Heeger (2009) model (Bortone, Olsen, & Scanziani, 2014; Olsen, Bortone, Adesnik, & Scanziani, 2012). Furthermore, the 6CT neurons also project to the thalamic reticular nucleus (TRN), which can contribute a surround-inhibition contrast-enhancing effect on top of the incoming attentional signal from the cortex (TRNCites).

- The temporal context information in the deep network allows the system to perform local predictive auto-encoder learning, where a condensed representation of the current state of the network is predicted on the basis of the temporal context representation from the prior 100 msec window. The dynamics of the predictive auto-encoder learning drive local learning signals throughout the local area network, to shape learning in the superficial and deep layers. Furthermore, these dynamics propagate through the bidirectional connections of the superficial layer network to drive longer-range error-driven learning throughout the entire posterior cortical network.

- The deep / thalamic network elements that support the predictive auto-encoder are as follows. The TRC neurons play the role of the *visible* neurons in an auto-encoder (i.e., representing the input / output patterns that are being encoded) – they are strongly driven by the feedforward 5IB projections from the previous brain area, and these projections are sparse and likely non-plastic, producing a relatively fixed recoding of the activity from the previous area as the *target* activation for the next layer to predict. Because of their alpha-frequency bursting dynamic, this driving input from the 5IB neurons occurs phasically roughly every 100 msec, and it constitutes the *plus phase* of error-driven learning. The *minus phase* precedes this plus phase, and represents the predicted state of the TRC encoding, which is driven by the layer 6CT neurons that project down to the TRC neurons associated with a given area. This projection is dense, much weaker than the driving 5IB projection, and likely plastic (Sherman & Guillery, 2006). The layer 6CT neurons rely on integrated inputs from earlier 6 corticocortical (6CC) neurons and 5IB neurons, to generate this predicted minus-phase state. The difference between the predicted TRC state in the minus phase, and the actual target TRC driven by 5IB firing in the plus phase represents the residual error signal, and synapses throughout the local network learn based on this temporal difference error signal, based on the CHL-like dynamics as captured in the existing Leabra learning equations (O'Reilly, 1996; O'Reilly et al., 2012; OReilly et al., 2015). Because the TRC projections go back up into cortex and synapse onto layer 4 and layer 6, they are ideally situated to send the minus and plus phase visible state information to the superficial and deep networks.

The detailed flow of activation according to these principles is illustrated in Figure 2, showing stimulus information coming into area V1, proceeding up to area V2 via the superficial network, and back down to area V1, again in the superficial network. All of this is modulated by ongoing layer 6CT activations driven from prior network state. Toward the end of the first 100 msec alpha cycle, the 5IB bursting from area V1 then drives a plus phase activation state in the TRC neurons of V2, which then propagates up through the V2 superficial and deep networks. Meanwhile, the V2 5IB neurons fire and drive a wave of new activation through the rest of the V2 deep network (6CC to 6CT), corresponding to an update in the attentional and temporal context state information for the V2 layer. The same update also happens in V1. During the next alpha cycle, these updated deep network activation states are then continuously communicated by the 6CT neurons to the TRC neurons in the thalamus (with TRN contrast enhancement), resulting in a minus

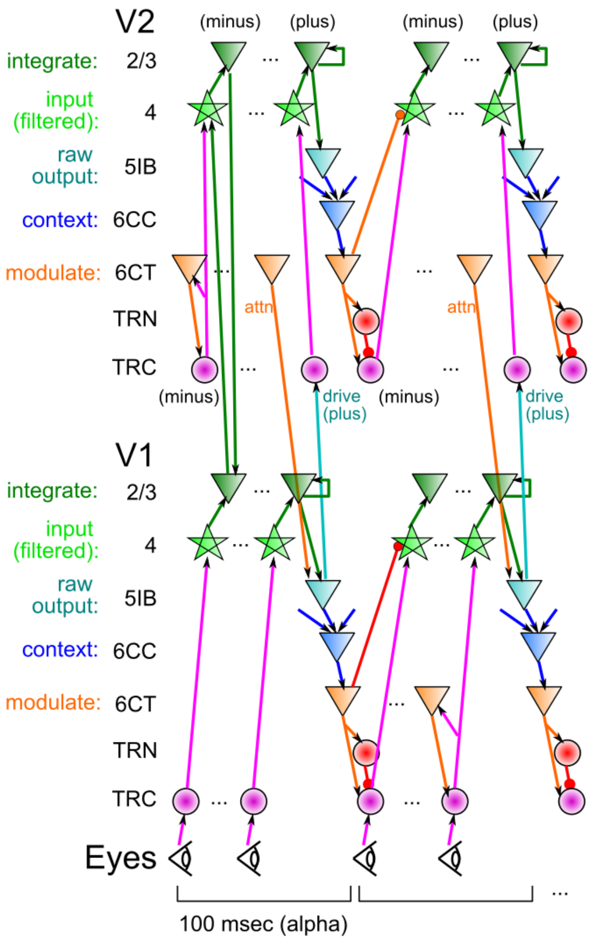

Figure 2: The temporal evolution of information flow in a DeepLeabra model across V1 and V2 layers. Information flows in from V1 to V2 and back in the superficial layer network (in green), while the deep network drives attentional modulation and temporal context information for predictive auto-encoder learning. Computationally, the superficial neurons serve to integrate information through constraint satisfaction, and they receive attentionally-filtered signals from layer 4 neurons that are modulated via 6CT (corticothalamic) projections. The 5IB neurons at the start of the deep layer network represent the most salient, raw (prior to contextual normalization) output of a given area, and the 6CC neurons integrate across this raw output to produce a context representation that supports the ability to make predictions about the next input state, and also drives renormalization of the 6CT attentional modulation signal.

phase expectation for the subsequent plus-phase state over the TRC's, and also a direct broadcast of the 6CT activation state up to the superficial cortex which serves to multiplicatively modulate the activation states there, producing an attentional modulation effect.

All of these dynamics end up working together to produce both dynamic, multiplicative attentional modulation and powerful error-driven predictive auto-encoder learning effects, both of which are critical computational advances over the existing Leabra model of the superficial network, which allow us to tackle the dynamics of something like the Raven's task using realistically-learned representations. The alpha cycle organizes both the learning and attentional update dynamics, in a synergistic fashion, with the deep / thalamic network providing an outer loop to the inner-loop of superficial layer constraint-satisfaction processing. This synergy includes the diffuse integrative *context* connections within the deep layer (e.g., supported by the 6CC neurons and other broad corticocortical connectivity among these deep neurons; Thomson, 2010; Thomson & Lamy, 2007), which are important for both the attentional and temporal context functions. For attention, the context information is essential for driving appropriate neighborhood-based inhibitory surround effects (Reynolds & Heeger, 2009), while for predictive learning the context is essential for capturing the prior state information in a form that then supports the ability to make accurate predictions about the next state. Computationally, this context information is very similar to that present in the simple recurrent network (SRN) framework (Elman, 1990, 1991; Jordan, 1989), as we elaborate below.

This framework gives a specific computational function to each of the major neuron types in the corticothalamic circuit (Figure 2), which goes beyond existing attempts to understand the differential functions of these neurons based on purely anatomical and physiological data (e.g., Thomson, 2010; Thomson & Lamy, 2007; Douglas & Martin, 2004; Markov et al., 2014). A major challenge for these data-based attempts to infer differential functionality is that the basic response properties of these neurons can appear quite similar in terms of the kinds of stimulus probes typically used in electrophysiological experiments (and in terms of what we observe in our computational models), despite each neuron type supporting a distinct computational function due to more subtle timing, response, and connectivity properties. Specifically, our model suggests that the superficial 2/3 neurons serve to integrate information through constraint satisfaction, based on attentionally-filtered input signals from layer 4 neurons that are modulated via 6CT (corticothalamic) projections. The 5IB neurons at the start of the deep layer network represent the most salient output of a given area, and the 6CC neurons serve to integrate across this output to produce a context representation that supports the ability to make predictions about the next input state, and also drives renormalization of the 6CT attentional modulation signal.

In subsequent sections, we elaborate a few of the major features of this model in more detail, and then proceed with a number of simulation tests of the model establishing its overall fit to relevant data in a number of domains, including attention, predictive learning, object recognition learning, and central elements of the Raven's task.

Attentional Dynamics

Our framework captures the same computations as the abstract Reynolds and Heeger (2009) model of attentional modulation. This model showed how seemingly different effects of attention (contrast gain vs. response gain) can both emerge out of a single unified framework, which responds differently as a function of task factors that shift it from exhibiting contrast gain vs. response gain effects. The model has 3 main steps to its computation: 1) the stimulus input is multiplied by the top-down attentional control weights, which results only in an *increase* for the parts of the input that are receiving attention – the rest of the attentional field is set to 1 and thus does not alter the unattended locations; 2) the resulting activation pattern is subject to a neighborhood-based pooling or integration over locations and orientations; 3) these pooled values then serve as a suppressive field that normalizes the net attentional modulation applied to the stimulus input, with the key overall result that now the unattended locations are actually suppressed in addition to the attended locations being enhanced. The relative balance of the enhancing vs. suppressive effects can shift depending on the relative sizes of the attentional spotlight and the stimulus input, producing the shift from contrast gain to response gain effects of attention.

Figure 3 shows how our model captures the essential computations of the Reynolds and Heeger (2009) model in different parts of the superficial and deep layer circuits. Working backward from the 6CT modulatory layer, we posit that this layer encodes a final normalized attentional mask that has an overall multiplicative or gain-field effect on neural activations in the superficial network, which is consistent with relevant data (Bortone et al., 2014; Olsen et al., 2012). Thus, where activations are strong in this layer, the corresponding superficial layer activations will remain strong, but where they are weaker, the superficial layer activations will be reduced. The normalization in 6CT occurs via inhibitory feedback circuits, both locally within layer 6 and through the TRN and TRC circuits of the thalamus (which then feed back into 6CT as well). This normalization process is affected by the 6CC layer prior to 6CT, which does the pooled integration over

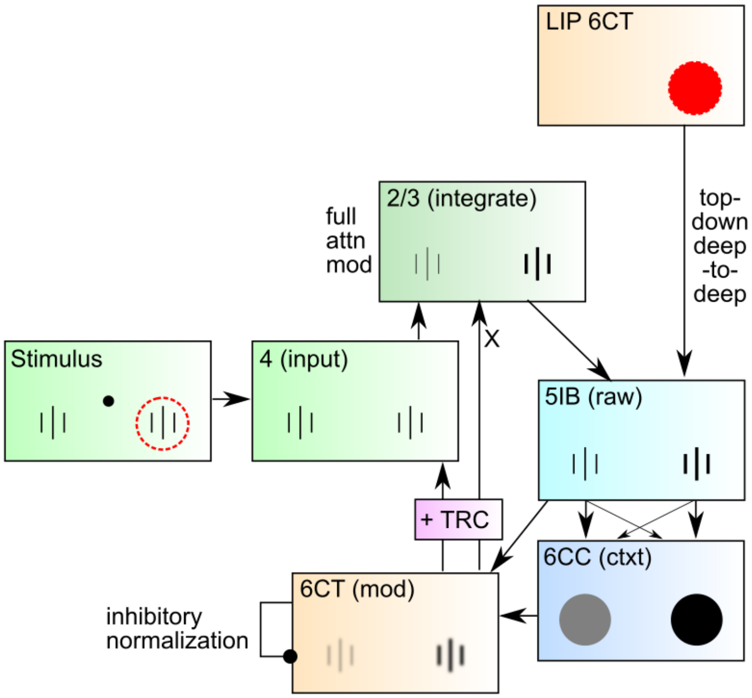

Figure 3: How attentional modulation is computed across the deep layers in response to a top-down attentional focus (as encoded in LIP of parietal cortex). Layer 4 receives bottom-up sensory input (initially equally weighted), which then drives superficial layers (2/3), which initially do not reflect the attentional modulation (not shown). The deep 5IB neurons integrate deep-to-deep top-down attentional inputs from LIP plus the local stimulus features from 2/3, to produce the *raw* deep output, prior to the contextual normalization process. The 6CC neurons integrate across the 5IB activations (context integration). 6CT then integrates this contextual and direct activation from 5IB, to produce, for the first time in the circuit, a properly renormalized multiplicative gain-field activation pattern, with surround inhibition both within the 6CT layer and further downstream in the TRN and TRC circuit providing the critical renormalization process. These 6CT activations then modulate (multiply) the superficial-layer activations to produce *both* an increase the attended location, and a decrease for the unattended location, as shown. In the biology, this modulation affects the layer 4 inputs (not shown) as well as 2/3. Our model subsumes layer 4 into layer 2/3 neurons.

space and features, and then feeds into 6CT. One step prior, area 5IB combines local stimulus features and the top-down attentional inputs from higher-level areas (e.g., LIP in this case, which has been shown to support spatially-organized attentional activations; Bisley & Goldberg, 2010). Thus, all of the same essential computations that are present in the Reynolds and Heeger (2009) model are distributed across these different deep layers.

The *folded-feedback* model of Grossberg (1999) (see Raizada & Grossberg, 2003 for a more elaborated version) also posits this same kind of attentional modulation dynamic between layer 6 and the superficial layers. Interestingly, top-down attentional signals, like those coming from LIP down to lower-level visual pathways, are preferentially communicated via a network of deep-to-deep projections (Figure 4; Markov et al., 2014). This network thus provides a very direct route (with direct shortcuts avoiding intervening layers in the hierarchy) for higher-level attentional signals to modulate lower-levels of the network, and is strongly consistent with the central role for the deep layers in supporting an attentional modulation function.

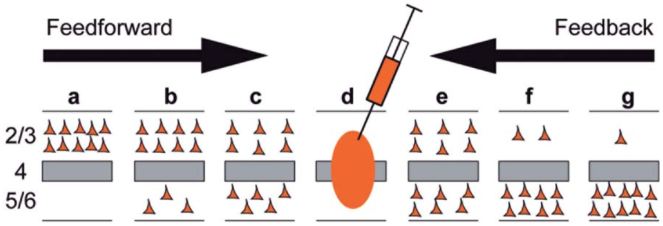

Figure 4: Overall distribution of the laminar sources of projections into a given cortical area (d, which is injected with retrograde label), showing that short-cut top-down projections (that skip intervening areas in the hierarchy) are much more likely to originate from deep layers (5/6), while short-cut feed-forward connections preferentially originate from superficial layers (2/3). Projections from immediately neighboring areas are balanced between superficial and deep. These projections also favor the corresponding target layers (i.e., superficial to superficial, deep to deep). Thus, long-range top-down projections directly influence the attentional deep network, while long-range feedforward projections preferentially communicate the integrative outputs from the superficial layers. Figure from Markov et al., 2014 based on their extensive analysis of connectivity data.

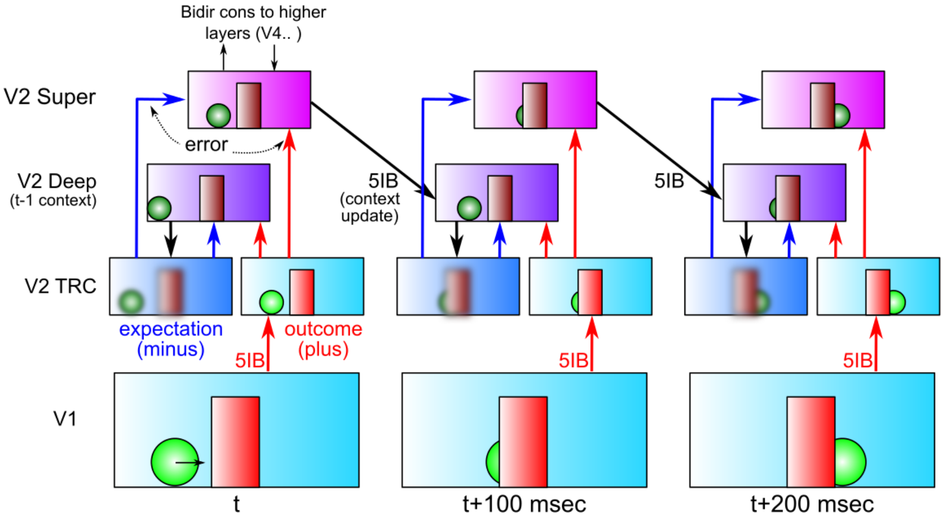

Figure 5: The temporal evolution of information flow in a DeepLeabra model predicting visual sequences, over a period of three alpha cycles of 100 msec each. The Deep network uses the prior 100 msec of context information to generate a prediction or expectation (minus phase) over the TRC units of what will come in next via the 5IB strong driver inputs from V1, which herald the next plus or target phase of learning. Error-driven learning occurs as a function of the temporal difference between the plus and minus activation states, in both superficial and deep networks, via the TRC projections into these networks. The 5IB bursting in V2 drives an update of the local temporal context information in V2, which is used in generating the minus phase in the next alpha cycle, and so on. These same 5IB cells drive a plus phase in higher layer TRC's as well, which perform the same kind of *local* predictive auto-encoder learning as shown for V2 here. This system is a predictive auto-encoder (generative model), because it is learning to generate a representation of the V1 inputs (as transformed via the relatively fixed V1 5IB to V2 TRC projection).

Predictive Auto-encoder Learning

The overall scheme for how predictive auto-encoder learning takes place in our framework is shown in Figure 5. The 5IB bursting activation coming from the lower layer (e.g., V1 for V2) defines the plus phase learning signal in the TRC neurons, which is implicitly compared against the immediately preceding prediction or expectation minus phase that is produced by top-down projections from 6CT neurons in the deep network to these TRC neurons. The TRC projections back up to layers 4 and 6 convey this plus – minus error signal difference over time to the superficial and deep networks, causing them to learn to better predict and efficiently encode the structure of visual information represented over time in the V2 TRC layer in the plus phases. Because this same network structure is replicated throughout the neocortex, we think that this same predictive auto-encoder learning is taking place everywhere, providing *locally-generated* error signals throughout the neocortex. Computationally, this is important because distally-generated error signals suffer from an exponential decay in strength as they propagate through network layers, causing intermediate layers in deep networks to have very weak learning signals. Thus, the local predictive auto-encoding dynamics, which are anchored by the strong feedforward driver signals carried by the 5IB neurons, can provide more robust learning signals in these intermediate layers. Furthermore, due to the auto-encoder nature of the system, these learning signals do not require any additional external learning signal – they simply use the next bottom-up sensory-driven input as the "teaching" signal. Finally, because of the extensive bidirectional connectivity between areas in the superficial and deep networks, error signals that arise locally in any given area also propagate throughout the network (albeit with the exponential decay factor).

Predictive auto-encoder learning has been explored in various frameworks, but the most relevant to our model comes from the application of the simple recurrent network (SRN) (Elman, 1990, 1991; Jordan, 1989), which employs the *simple* trick of copying the current internal (hidden) layer representation to a context layer that then acts as an additional input to the hidden layer for generating a prediction of what will happen on the next time step. In effect, we hypothesize that the time step for updating an SRN-like context layer is the 100 msec alpha cycle, and during a single alpha cycle, considerable bidirectional constraint satisfaction neural processing is taking place within a DeepLeabra network. This contrasts with the standard SRN, which is typically implemented in a feedforward backpropagation network, where each time step and context update corresponds to a single feedforward activation pass through the network. Despite this important difference, and several others that we discuss below, there are some critical computational lessons that we adopt directly from the SRN.

One of the most powerful features of the SRN is that it enables error-driven learning, instead of arbitrary parameter settings, to determine how prior information is integrated with new information. Thus, SRNs can learn to hold onto some important information for a relatively long interval, while rapidly updating other information that is only relevant for a shorter duration (e.g., Cleeremans, Servan-Schreiber, & McClelland, 1989; Cleeremans, 1993). This same flexibility is present in our DeepLeabra model. Furthermore, because this temporal context information is hypothesized to be present throughout the entire neocortex (in every microcolumn of tissue), the DeepLeabra model provides a more pervasive and interconnected form of temporal integration compared to the SRN, which typically just has a single temporal context layer associated with the internal "hidden" layer of processing units.

An extensive computational analysis of what makes the SRN work as well as it does, and explorations of a range of possible alternative frameworks, has led us to an important general principle: *future outcomes determine what is relevant from the past.* At some level, this may seem obvious, but it has significant implications for predictive learning mechanisms based on temporal context. It means that the information encoded in a temporal context representation cannot be learned at the time when that information is presently active. Instead, the relevant contextual information is learned on the basis of what happens next. This explains the peculiar power of the otherwise strange property of the SRN: the temporal context information is preserved as a *direct copy* of the state of the hidden layer units on the previous time step (Figure 6), and then learned synaptic weights integrate that copied context information into the next hidden state (which is then copied to the context again, and so on). This enables the error-driven learning taking place in the *current* time step to determine how context information from the *previous* time step is integrated. And the simple direct copy operation eschews any attempt to shape this temporal context itself, instead relying on the learning pressure that shapes the hidden layer representations to also shape the context representations. In other words, this copy operation is essential, because there is no other viable source of learning signals to shape the nature of the context representation itself (because these learning signals require future outcomes, which are by definition only available later).

The direct copy operation of the SRN is however seemingly problematic from a biological perspective:

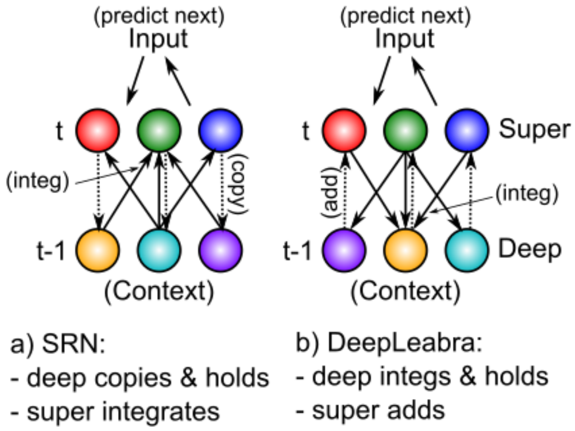

Figure 6: How the DeepLeabra temporal context computation compares to the SRN mathematically. **a)** In a standard SRN, the context (deep layer biologically) is a copy of the hidden activations from the prior time step, and these are held constant while the hidden layer (superficial) units integrate the context through learned synaptic weights. **b)** In DeepLeabra, the deep layer performs the weighted integration of the soon-to-be context information from the superficial layer, and then holds this integrated value, and feeds it back as an additive net-input like signal to the superficial layer. The context net input is pre-computed, instead of having to compute this same value over and over again. This is more efficient, and more compatible with the diffuse interconnections among the deep layer neurons. Layer 6 projections to the thalamus and back recirculate this pre-computed net input value into the superficial layers (via layer 4), and back into itself to support maintenance of the held value.

how could neurons copy activations from another set of neurons at some discrete point in time, and then hold onto those copied values for a duration of 100 msec, which is a reasonably long period of time in neural terms (e.g., a rapidly firing cortical neuron fires at around 100 Hz, meaning that it will fire 10 times within that context frame). However, there is an important transformation of the SRN context computation, which is more biologically plausible, and compatible with the structure of the deep network (Figure 6). Specifically, instead of copying an entire set of activation states, the context activations (generated by the phasic 5IB burst) are immediately sent through the adaptive synaptic weights that integrate this information, which we think occurs in the 6CC (corticortical) and other lateral integrative connections from 5IB neurons into the rest of the deep network (Thomson, 2010; Thomson & Lamy, 2007). The result is a *pre-computed net input* from the context onto a given hidden unit (in the original SRN terminology), not the raw context information itself. Computationally, and metabolically, this is a much more efficient mechanism, because the context is, by definition, unchanging over the 100 msec alpha cycle, and thus it makes more sense to pre-compute the synaptic integration, rather than repeatedly re-computing this same synaptic integration over and over again.

There are a couple of remaining challenges for this transformation of the SRN. First, the pre-computed net input from the context must somehow persist over the subsequent 100 msec period of the alpha cycle. We hypothesize that this can occur via NMDA and mGluR channels that can easily produce sustained excitatory currents over this time frame. Furthermore, the reciprocal excitatory connectivity from 6CT to TRC and back to 6CT could help to sustain the initial temporal context signal. Second, these contextual integration synapses require a different form of learning algorithm that uses the sending activation from the prior 100 msec — there are cases where such a temporal offset in learning has been documented (cerebellum-ltd-delay), and biophysically the time constants in the relevant calcium and second messenger pathways involved in synaptic plasticity could plausibly accommodate this amount of temporal offset.

Finally, we note that the use of the TRC units as the effective *visible* units in our predictive auto-encoder scheme is only one of various possible ways of conceptualizing the arrangement of such a system within the thalamocortical networks, and we had proposed a different configuration previously, which we referred to as *LeabraTI* (temporal integration) (Kachergis, Wyatte, O'Reilly, de Kleijn, & Hommel, 2014). The current configuration has many advantages over the previous one, because of the unique position of the TRC neurons within the overall network. The TRC neurons receive two major inputs: the strong driver inputs from the 5IB cells of the previous layer, and the weaker, far more numerous top-down inputs from the same layer that they project to (Sherman & Guillery, 2006). The phasic alpha-frequency bursting dynamics of the 5IB driver inputs naturally provides a distinction between a *clamped* plus phase-like target activation state, versus a more labile state driven only by the weaker top-down inputs. Note that although Sherman and Guillery (2006) characterize these top-down inputs as modulatory, *in vivo* electrophysiological recording data shows constant steady activation of TRC neurons across multiple alpha trials worth of time, suggesting that these top-down projections are capable of driving TRC activation in between the 5IB bursting (Bender, 1982; Petersen, Robinson, & Keys, 1985; Bender & Youakim, 2001; Robinson, 1993). The LeabraTI model hypothesized that higher layers attempted to reconstruct the activation states over the superficial layers of the layers below them, which raised many problems having to do with creating a plausible (and computationally effective) difference between the minus and plus phase states of these layers. Thus, the configuration of the TRC neurons within the overall network seems suspiciously ideal for their use as a substrate for predictive auto-encoder learning. Furthermore, using a single layer driven bidirectionally for the visible layer neurons as we do with the TRC neurons is much more efficient and natural than the two separate layers (input and output) that are required in the typical feedforward SRN framework.