Deep Predictive Learning in Vision

Randall C. O'Reilly (randy.oreilly@colorado.edu)
Department of Psychology and Neuroscience, University of Colorado Boulder

John Rohrlich (john.rohrlich@colorado.edu)
Department of Psychology and Neuroscience, University of Colorado Boulder

Abstract

How does the neocortex learn and develop the foundations of our high-level cognitive abilities? We present a comprehensive framework spanning biological, computational, and cognitive levels, providing a coherent answer supported by data. Learning is based on making predictions about what the senses will report at 100 msec (alpha frequency) intervals, and adapting synaptic weights to improve prediction. The pulvinar nucleus of the thalamus serves as a projection screen upon which predictions are generated, through deep-layer 6 corticothalamic inputs from multiple brain areas. The bottom-up, sparse, driving inputs from layer 5 intrinsic bursting neurons provide the target signal, and the temporal difference between it and the prediction reverberates throughout cortex, driving synaptic changes that approximate error backpropagation, using only local activation signals in equations derived from a detailed biophysical model. We test this framework of unsupervised predictive learning with a model of the visual system that incorporates two central principles: top-down input from compact, high-level, abstract representations is required for accurate prediction of low-level sensory inputs; and the collective, low-level prediction error is progressively partitioned to enable extraction of separable factors that drive learning of high-level abstractions. Our model self-organized invariant object representations of 100 objects from simple movies and accounts for a wide range of data.

Keywords: neocortex; predictive learning; pulvinar; vision

Introduction

What is the nature of the remarkable neocortical learning and maturational mechanisms that result in the development of our considerable perceptual and cognitive abilities? In other words, where does our knowledge come from? Phenomenologically, it appears to magically emerge after several months of gaping at the world passing by — what is the magic recipe for extracting high-level knowledge from an ongoing stream of perceptual experience? Answering this question has been the ultimate goal of many lines of research, at many levels of analysis from synapses to machine learning algorithms and psychological theories. Despite many advances at each of these levels of analysis, we still lack an overall framework providing a comprehensive answer to this question. Here we propose such a framework, one that provides a broad and deep integration of many different sources of data. This biologically grounded framework is implemented in a computer model that demonstrates both its computational function and its ability to account for a wide range of data.

Our core hypothesis, also advanced by other researchers going back at least to Helmholtz in 1867 (von Helmholtz, 2013), is that learning can emerge from the largely passive sensory experience of babies because each moment is an opportunity for predictive learning. Underlying the seemingly passive behavior is an active neural network generating predictions for what will happen next, and a process that drives learning from the differences between these predictions and what actually does occur. Within this general framework, several natural questions arise: How frequently are predictions generated and what stimulates their generation? How exactly are the predictions compared with reality, and what form does that critical difference (i.e., the *prediction error*) take, so that it can drive learning? And how can the brain simultaneously represent both a prediction and the sensory ground truth, without getting them mixed up?

Specific Hypotheses

Our specific hypotheses are as follows: Predictions (in sensory posterior cortex, at least) are generated every 100 msec (i.e., the alpha rhythm), driven fundamentally by deep layer 5IB intrinsic-bursting neurons which burst at this frequency, entrained via circuits interconnecting the neocortical deep layers with the higher-order sensory thalamus (the *pulvinar*). We view this as a subconscious process specifically for sensory predictive learning — other time scales and forms of predictive learning may occur in other brain areas. These predictions are generated within the deep neocortical layers (5 and 6), based on time-delayed information from the prior 100 msec, and projected broadly onto the pulvinar thalamic relay cells, via the numerous, weaker "top-down" pathway from neocortical layer 6 (Sherman & Guillery, 2006). After about 75 msec of reflecting these top-down

predictions, the sensory bottom-up ground truth drives the pulvinar, via very sparse, strong projections from the 5IB neurons in lower cortical areas (Sherman & Guillery, 2006), and this *temporal difference* reflects the prediction error signal. Thus, unlike most other predictive / generative learning frameworks, we do not propose a population of neurons whose activation explicitly reflects the prediction error — instead the error is implicitly reflected in the temporal dynamics of activation signals emanating from the pulvinar.

The pulvinar projects broadly throughout the posterior cortex, and this temporal difference at the alpha frequency can drive learning throughout the cortex to improve the accuracy of the predictions generated by the deep layers. Furthermore, while the deep layers are driving their predictions, the superficial neocortical layers are integrating bottom-up and top-down information about the current state of both the environment and the organism, and also learning to improve these representations via the same temporal-difference prediction error signal. Thus, we propose a clear anatomical separation between the predictive (deep layers) and current-time (superficial layers) representations in the cortex — every alpha cycle, the superficial layer state provides the input to the deep layers (again via layer 5IB bursting) that will be used in generating the predictions for the next alpha cycle.

Predictive Learning to Develop Invariant Object Representations

As for the question of how far predictive learning can go, we focus on the widely-studied domain of invariant representations of objects. Such representations are widely recognized as having great adaptive value to an organism, and form the foundation of much of our semantic understanding of the world. However, to develop these representations models typically require training with explicit, invariant category labels. If predictive learning can be shown to form such representations in a purely unsupervised manner (i.e., strictly through the process of predicting subsequent sensory inputs, without any additional high-level category information), then it seems more likely that predictive learning could support a reasonably wide range of higher-level cognitive learning. We explore this question in the context of a simplified, analytic environment where one out of 100 different possible patterns moves in a random direction (or remains still) while the model makes random saccades every 200 msec. This captures the most basic aspects of the visual world: patterns (objects) that are generally stable over time but follow Newton's first law of motion, while also incorporating eye movements, which are the main reliable form of motor control available to a baby. We find that indeed invariant representations do form in the highest layers in the model, corresponding to those in inferotemporal cortex (IT) in the primate brain, and demonstrate that these representations play a critical role in the overall predictive learning process by compactly and stably encoding the visual features present in objects even as the spatial locations where those features appear changes.

A Hierarchical Generative Model

Computationally, our framework is a form of a *hierarchical generative model*, which have been widely explored as models of brain / cognitive function (and we restrict our discussion to that subset, broadly defined, as opposed to the broader machine learning field). These models are typically trained progressively from the bottom-up (i.e., layer-by-layer), and according to a relatively strict hierarchy where each layer learns to predict the behavior of the layer below it. We found this approach to have significant limitations, and instead discovered two critical principles that were necessary for the development of systematic, high-level, abstract knowledge representations in our model: 1. Compact, high-level abstract representations are essential for accurate prediction generation at the lowest levels, and thus there must be extensive top-down short-cut projections from the highest levels of the hierarchy down to the lowest levels; and 2. The overall prediction error (broadcast by the pulvinar as a temporal difference) must be progressively and opportunistically partitioned by differentially-specialized such high-level pathways, with simpler factors learned earlier and thus factored out from the overall predictive error signal, thereby concentrating the efficacy of the remaining signal for learning the other aspects of the overall prediction problem.

Although many generative models are discussed in terms of generating predictions, many of them do not actually include an explicit temporal divide, and instead end up learning by reconstructing the *current* sensory input (e.g., an *auto-encoder* in neural network terms). These kinds of auto-encoders require various constraints to avoid degenerate solutions, and it remains unclear whether such models can produce systematic abstract internal representations in a purely self-organizing manner (typically they are subsequently trained with standard explicit object category labels, for example). By contrast, the task of predicting the *future* sensory input avoids many of these problems because, as the saying goes, prediction is difficult, *especially about the future*.

The DeepLeabra Predictive Learning Framework

In recognition of the critical predictive role of deep neocortical layers, and the ability to train deep hierarchi-

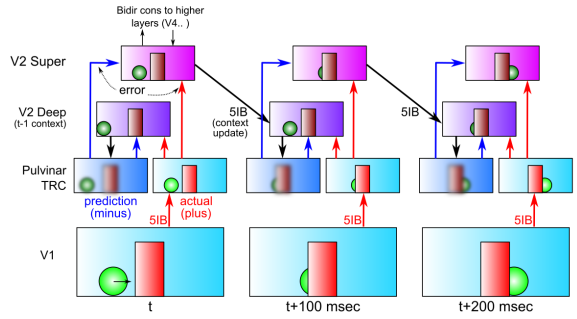

Figure 1: Schematic illustration of the temporal evolution of information flow in a DeepLeabra model predicting visual sequences, over a period of three alpha cycles of 100 msec each.

cal networks, we refer to our computational model as the *DeepLeabra* learning algorithm, building on our earlier *Leabra* mechanism that performed the same temporal-difference-based error-driven learning in bidirectionally-connected networks, but previously based only on the superficial layers of the neocortex (O'Reilly, Hazy, & Herd, 2016; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012; O'Reilly & Munakata, 2000; O'Reilly, 1996).

Figure 1 provides an overall schematic for how predictive learning takes place in our framework, showing area V2 predicting the next pattern of activation on V1, over the period of three alpha-cycle "movie frames". The V2 deep-layer neurons drive activation of a minus-phase prediction over the pulvinar, and then in the plus phase the 5IB neurons in area V1 drive the pulvinar with the actual sensory input state, and the temporal difference between the two represents the error signal that trains the superficial and deep layers of V2 to create better representations for making a more accurate prediction next time around. This same cycle of prediction and training occurs for all the layers of the visual system.

The neocortex in our model is composed of two separable but tightly interacting sub-networks, the superficial and the deep / thalamic (pulvinar). 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 area, and feedback coming from 2/3 in one area back to 2/3 in an earlier area; Rockland & Pandya, 1979; Felleman & Van Essen, 1991; Markov et al., 2014). The superficial network represents the current state of the environment and internal state of the organism, at multiple different levels of abstraction, all mutually interacting. It can be described computationally in terms of a classic Hopfield network / Boltzmann machine constraint satisfaction system (Hopfield, 1982; Ackley, Hinton, & Sejnowski, 1985).

The deep / thalamic network starts in each area with the layer 5b intrinsic bursting (IB) neurons (Connors, Gutnick, & Prince, 1982; Sherman & Guillery, 2006; Franceschetti, Guatteo, Panzica, Sancini, Wanke, & Avanzini, 1995; Flint & Connors, 1996), which receive inputs from local superficial neurons and top-down projections from other areas. 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-area thalamic areas. The deep / thalamic network in the posterior cortex is directly responsible for generating predictions over the pulvinar. It must be phasically shielded from the current state information in the superficial layers, to be forced to generate a prediction, as opposed to simply copying the current input state (in which case it would become a simple auto-encoder).

The brief, phasic bursting of the 5IB neurons is the essential mechanism in our model that ensures that bottom-up, current-state information only penetrates the deep layers phasically, not continuously, thus enabling true predictions to be generated. During the minus phase, when it is generating the next prediction, the deep state reflecting the last 5IB burst of activity is sustained and elaborated through regular spiking layer 6 neurons (i.e., layer 6CT corticothalamic neurons; Thomson, 2010) that project to the thalamic relay cells (TRC) of the pulvinar, which then project back to these same 6CT neurons (and up to the layer 4 inputs to the superficial network).

Testing the Framework

To test the above predictive learning mechanisms, we applied it to a simple visual prediction task with short "movies" of objects undergoing constant self-motion, and randomly directed saccades with an efferent copy of the upcoming saccade motor plan. After the first frame of such a movie the subsequent frames should be fully predictable, so our first test was whether the model could learn to accurately predict these subsequent frames. We were also interested in the extent to which these same predictive learning mechanisms could develop high-level abstract representations of objects that can then provide a more systematic basis for intelligent behavior. For example, by developing invariant object representations, an organism would be able to systematically respond appropriately to the presence of objects regardless of the perceptual details in which that object was viewed. Figure 2, shows decoding accuracy improving in higher visual cortical layers without any supervised learning.

Overall, we found a strong correspondence between the successful principles for improving overall network performance, and known features of the biology. The ex-

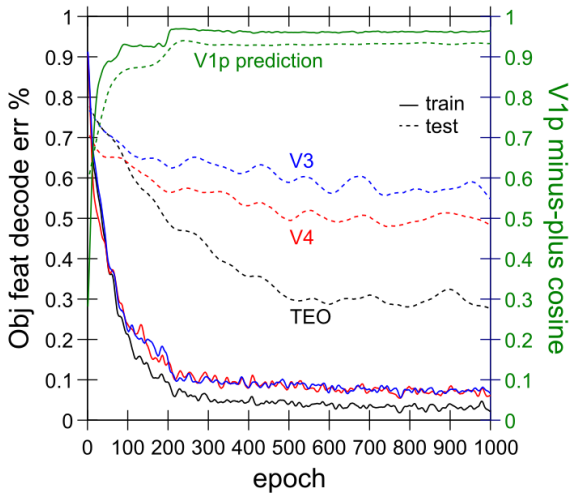

Figure 2: Learning curves for full model, showing accuracy (proportion error) in decoding the object features from each of 3 different layers (V3, V4, TEO), and overall prediction accuracy in terms of minus vs. plus phase cosine over the V1p pulvinar layer, at trial 3 (the last trial), which is nearly perfect. Notice that TEO has developed much more systematic object representations than other layers.

tent and depth of this correspondence suggests that structural and developmental properties of the mammalian visual neocortex may have evolved to support the same kinds of computational principles of predictive learning.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*(1), 147–169.

Connors, B. W., Gutnick, M. J., & Prince, D. A. (1982). Electrophysiological properties of neocortical neurons in vitro. *Journal of Neurophysiology*, *48*(6), 1302–1320.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, *1*(1), 1–47.

Flint, A. C., & Connors, B. W. (1996). Two types of network oscillations in neocortex mediated by distinct glutamate receptor subtypes and neuronal populations. *Journal of Neurophysiology*, *75*(2), 951–957.

Franceschetti, S., Guatteo, E., Panzica, F., Sancini, G., Wanke, E., & Avanzini, G. (1995). Ionic mechanisms underlying burst firing in pyramidal neurons: Intracellular study in rat sensorimotor cortex. *Brain Research*, *696*(1–2), 127–139.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., Lamy, C., Misery, P., Giroud, P., Ullman, S., Barone, P., Dehay, C., Knoblauch, K., & Kennedy, H. (2014). Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex: Cortical counterstreams. *Journal of Comparative Neurology*, *522*(1), 225–259.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2016). The Leabra cognitive architecture: How to play 20 principles with nature and win! In S. Chipman (Ed.), *Oxford handbook of cognitive science*. Oxford University Press.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu.

Rockland, K. S., & Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, *179*(1), 3–20.

Sherman, S., & Guillery, R. (2006). *Exploring the Thalamus and Its Role in Cortical Function*. Cambridge, MA: MIT Press.

Thomson, A. M. (2010). Neocortical layer 6, a review. *Frontiers in Neuroanatomy*, *4*(13).

von Helmholtz, H. (2013). *Treatise on Physiological Optics, Vol III*. Courier Corporation.