Deep Predictive Learning as a Model of Human Learning

Randall C. O'Reilly, Jacob L. Russin, and John Rohrlich
Departments of Psychology and Computer Science
Center for Neuroscience
University of California, Davis
1544 Newton Ct
Davis, CA 95616
oreilly@ucdavis.edu

September 17, 2019

Abstract:

Now that's abstract..

The success of deep convolutional neural networks (DCNN's; cites) in object recognition and many other domains raises the question of how well they model human learning, at both neural and cognitive levels. Although the engine of these models, error backpropagation (Rumelhart, Hinton, & Williams, 1986), has long been questioned on biological grounds (Crick, 1989), many biologically-plausible mechanisms have been proposed (O'Reilly, 1996etc). However, the need for massive amounts of labeled data still makes these models cognitively implausible: non-human primates and infants learn to recognize and categorize objects without massive amounts of labeled data (cites). A promising alternative, biologically-plausible approach is to use predictive error-driven learning, where error signals arise from differences between a prediction of what will happen next, and what actually does happen (Elman, 1990). Here we show that distinctive circuits in cortex and thalamus have several properties that specifically support predictive error-driven learning. Furthermore, when implemented in a computational model employing a biologically-plausible form of error backpropagation (O'Reilly, 1996; O'Reilly & Munakata, 2000; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012), along with several other important properties of the mammalian visual system, the model learns to systematically categorize 3D objects according to invariant shape properties, in a way that matches human judgments of these same objects, and is consistent with neural representations in inferotemporal (IT) cortex in humans and other primates. Comparison models with the same structure but based on standard non-biological error-backpropagation learning demonstrate the importance of predictive learning for developing categorical encodings of any form, but the binary categories formed by these models do not go much beyond the similarities present at the lowest visual levels, and do not fit human judgments. Thus, incorporating biological properties of the brain can potentially provide a better understanding of human learning at multiple levels relative to existing DCCN models.

Motivated by biological evidence, we hypothesize that sensory predictions in posterior neocortex are generated roughly every 100 msec (i.e., the alpha rhythm, 10 Hz), by neurons in the deep layers of the neocortex that project to the pulvinar nucleus of the thalamus (Figure 1). The pulvinar represents this top-down

R. C. O'Reilly is Chief Scientist at eCortex, Inc., which may derive indirect benefit from the work presented here.

Supported by: ONR grants N00014-18-1-2116 (Deep learn), N00014-14-1-0670 / N00014-16-1-2128 (Bidir vis), N00014-18-C-2067 (MDM)

This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research.

prediction for roughly 75 msec of the alpha cycle, after which point the layer 5IB intrinsic-bursting neurons send strong, bottom-up driving input to the pulvinar, representing the actual sensory stimulus (Sherman & Guillery, 2006). These 5IB neurons burst at the alpha frequency, determining the overall timing of the predictive learning cycle, along with other dynamic parameters of the thalamocortical circuit (cites). Consistent with the temporal-difference nature of biologically-plausible error signals in this framework, the prediction error is implicit in the temporal difference between these two periods of activity within the alpha cycle over the pulvinar. The pulvinar sends broad projections back up to all of the areas that drive top-down predictions into it (Shipp, 2003; Mumford, 1991), thus broadcasting this error signal to drive local synaptic plasticity in the neocortex. This is mathematically equivalent to performing gradient descent to minimize overall prediction errors. This computational framework makes sense of otherwise puzzling anatomical and physiological properties of the cortical and thalamic networks (Sherman & Guillery, 2006), and is consistent with a wide range of detailed neural and behavioral data regarding the effects of the alpha rhythm on learning and perception (many cites). It has many testable differences from other existing forms of predictive learning that have been proposed over the years, at varying levels of biological detail (Friston, Rao, etc).

Figure 1a: 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. During each alpha cycle, the V2 Deep layer uses the prior 100 msec of context information to generate a prediction or expectation (minus phase) over the pulvinar thalamic relay cell (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 area 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 encoded via the relatively fixed V1 5IB to pulvinar projection).

b: computational model that learns by predicting what will be seen next..

TODO c: show frames of movie

A critical question for predictive learning is whether it can develop higher-level, more abstract ways of representing sensory inputs, that go beyond the raw low-level structure of these inputs, while learning on nothing other than predicting these raw visual inputs. For example, can predictive learning really replace the labor-intensive process of generating human-labeled image datasets where abstract category information is explicitly used to train object recognition models via error-backpropagation? From a cognitive perspective, there is considerable evidence that non-verbal primates, and pre-verbal human infants, naturally develop abstract categorical encodings of visual objects in IT cortex (cites), without relying on any explicit external categorical labels. Existing predictive-learning models based on error backpropagation have not demonstrated this result (cites); instead, these models are typically tested on their ability to subsequently train using the human-generated labels, but there are many ways in which prior training could facilitate such training without forming systematic categories de-novo.

To determine if our biologically-based predictive learning model can naturally form such categorical encodings in the complete absence of externally-provided category labels, we showed the model brief movies of 176 specific 3D object exemplars drawn from 20 different categories selected from the CU3D-100 dataset (cite). The objects moved and rotated in 3D space over 8 movie frames, where each frame was sampled at the 10 Hz alpha frequency (note that the minimum frame rate for actual movies (24 Hz) is just over the Nyquist 2x sampling frequency for 10hz alpha, which is further suggestive evidence for the idea that our visual system operates at this frequency). There were also saccadic eye movements every other frame, with an efferent copy signal to enable full prediction of the effects of the eye movement, which allows the model

to capture predictive remapping (a widely-studied signature of predictive learning in the brain; cites), and introduces further predictive-learning challenge. The only learning signal available to the model was the temporal difference prediction error between what it predicted to see in the next frame, compared to what was actually seen.

Fig 2a: Category similarity structure that developed in biologically-based predictive learning model, showing 1-cosine similarity for each object against every other object, organized by the broad shape categories formed in the highest IT layer of the model (TE). Blue cells have high similarity, and model has learned block-diagonal clusters or categories of high-similarity groupings, organized according to overall object shape.

Fig2b: human similarity ratings for the same objects. c) Macaque monkey category similarity structure (DiCarlo et al), for semantic categories: animals, cars, chairs, faces, fruits, planes, and tables. Although the categories are different, and monkeys have access to much more multimodal information than our model, the qualitative patterns are similar. d) Emergence of abstract category structure over the hierarchy of layers. Red line shows cosine similarity between the RSA item-similarity structure for TE (shown in panel a) vs. RSA for every other layer, and black line shows RSA for V1 layer vs. every other layer. Cosine = 1 means the structure is identical, and 0 = orthogonal. Both show that IT layers (TEO, TE) progressively differentiate from raw input similarity structure present in V1.

We performed a representational similarity analysis (RSA) on the learned activity patterns at each layer in the model, and found that the highest IT layer (TE) produced a systematic organization of the 3D objects into 5 different categories (identified post-hoc based on visual inspection), which visually correspond to the overall shape of the objects as shown in Figure 2a (pyramid-shaped, vertically-elongated, round, boxy / square, and horizontally-elongated). TODO: This organization of the objects matches that produced by humans making shape similarity judgments on the same set of objects (Fig 2b; see supplemental online material for methods and further analysis). Furthermore, this overall categorical structure is similar to the categorical nature of visual object representations in macaque monkeys (Fig 2b; DiCarlo et al), although different sets of objects were used (see supplemental material for further discussion).

Critically, Figure 2d shows that the overall RSA similarity structure present in IT layers (TEO, TE) of the biological model is significantly different from the similarity structure at the level of the V1 primary visual input layer of the model. Thus, the model has learned an abstract way of organizing these objects that goes beyond the raw pixel-level information, despite only being trained to generate accurate predictions about what the scene will look like next, at the level of this low-level visual representation. Furthermore, this abstract category organization reflects the overall visual shapes of the objects as judged by human participants looking at the same objects, suggesting that the model is learning about the same kind of overall geometrical shape information that is apparent once these objects are encoded in an invariant way that abstracts across the motion, rotation, and scaling transformations present in the V1 visual inputs. We further verified that at the highest IT levels in the model, a consistent representation is present across different views of the same object (supplementary Figure X), and, as shown in Figure 2a, across multiple objects within the same category (the dark blue cells there reflect close similarity across objects).

Fig 3a: Category similarity structure in the highest IT layer (TE) of the backpropagation model, which is essentially binary in structure. Some ground truth object categories (e.g., piano, slr camera) are split across these two categories, as evidenced by the transition not taking place cleanly along the black horizontal and vertical dividing lines (which split along the ground truth categories). Representations here are overall much more similar to each other colorscale is renormalized to use full color range compared to equivalent 2a figure. 3b: Comparison of V1 RSA similarity structure across layers, for backpropagation model (red line) compared to data previously shown in 2d for biological model (black line). Unlike the biological model, the V1-level structure persists largely intact throughout the backprop model, except in layer V3s which is most directly influenced by spatial prediction errors. Thus, there is no evidence that the backpropagation model

has developed significantly different structure beyond that present in the lowest V1 levels. 3c: Comparison of TE RSA similarity structure across layers, for all combinations of backpropagation vs. biological model (black line is same as 2d). The red line shows that, consistent with 3b, the TE-level structure in backprop is consistent with most layers except V3, again indicating a lack of significant novel structure in TE. The blue line shows that the TE structure in the biological model (2a) is not a good fit for any layers in the backprop model, and the green line shows that IT (TEO, TE) layers of the biological model strongly diverge from the backprop TE representations.

Figure 3 shows the results from a non-biological backpropagation-based version of the same model architecture. In this model, the highest layers in the network form a more degenerate binary category structure overall, and the detailed item-level similarity structure does not diverge significantly from that present at the lowest V1 inputs, indicating that it has not formed novel systematic structured representations, in contrast to those formed in the biologically-based model. Thus, it is clear that the additional biologically-motivated properties of the original model are playing a critical role in the development of abstract categorical representations. These properties include: excitatory bidirectional connections, inhibitory competition, and a Hebbian form of learning that serves as a regularizer in addition to the predictive error-driven learning (O'Reilly, 1998; O'Reilly & Munakata, 2000). Ongoing work is attempting to determine which of these is essential, perhaps all of them, by introducing some of these properties into the backpropagation model, though this is difficult because full bidirectional recurrent activity propagation, which is essential for conveying error signals top-down in the biological network, is incompatible with the standard efficient form of error backpropagation, and requires much more computationally intensive and unstable forms of fully recurrent backpropagation (Williams et al, etc). Furthermore, Hebbian learning requires inhibitory competition which is difficult to incorporate within the backpropagation framework results shown next demonstrate the importance of this Hebbian learning factor within the biological model.

Fig 4a: contributions of the dorsal where pathway in the biological model, including LIP (lateral inferior parietal sulcus area), V3, and DP (dorsal prelunate). Lesioning this pathway causes the IT (TEO, TE) representations to be much more similar to those in V1, and, likewise, the TE representations to be relatively similar to those throughout the network. Thus, having the spatial pathway regress out the spatial component of prediction error is essential for concentrating the residual error on object features, which are learned by the ventral what pathway into IT. b) shows that the what / where organization is also important in the backprop version of the model what little organization it developed that diverged from V1 is eliminated when the dorsal where pathway is fully lesioned. Furthermore, when trained as a pure auto-encoder of the current image (instead of predicting the next), it fails to develop anything other than a pure V1-level surface representation. 4b: hebbian param diff easily has big effects, key for connecting to bio stuff.

Despite these difficulties, the biological model results were based on extensive experimentation and theorizing about several other specific properties that should be important for enabling it to learn these abstract categorical representations, and we are able to directly test the relevance of many of these factors. For example, we hypothesized that separating the overall prediction problem between spatial "where" vs. non-spatial "what" pathways, a well-established property of the dorsal vs. ventral streams of the visual system (Ungerleider & Mishkin, 1982), would be essential for partitioning the overall prediction error. The spatial where pathway can learn relatively quickly to predict the overall spatial trajectory of the object (and anticipate the effects of saccades), and thus effectively regress out that component of the overall prediction error, leaving the residual error concentrated in object feature information, which can train the ventral what pathway to develop abstract visual categories. Figure 4a shows that, indeed, when the where pathway is lesioned to varying degrees, it systematically impairs the formation of the abstract categorical representations in the intact what pathway. Figure 4b shows the same manipulation on the backpropagation model, with similar results: what little deviation it exhibited from the pure V1 similarity structure is eliminated with a complete dorsal pathway lesion. In both models, lesioning the dorsal where pathway significantly impacted overall

prediction error, as would be expected (Supplemental Figure X). Regarding the other biological properties, Figure 4c shows the impact of changing Hebbian learning parameters, reducing the strength of lateral context connections within the IT pathway, and eliminating the temporally-delayed nature of these connections, all of which significantly impaired the formation of categorical representations. The effects of other tests are reported in the supplemental material.

A final experiment tested the importance of performing predictive learning, as compared to an auto-encoder, which learned to encode and reproduce the visual inputs within a single frame, as compared to having to predict the next frame from the current one. Auto-encoders are closely related to predictive learning models, and have been widely explored (cites). However, the encoder paradigm is more susceptible to degenerate solutions amounting to performing a mindless copy as opposed to a deep encoding, and considerable effort must be made to discourage such solutions, whereas prediction, being about the future, is not susceptible to these solutions (except in fully static environments where the two approaches converge). We tested an auto-encoder version of the backpropagation model which reproduced the current input frame, instead of trying to predict the next frame in the movie, and found that this completely eliminated any development of non-V1 similarity structure in the model (Figure 4b).

In conclusion, we have demonstrated that learning based strictly on predicting what will be seen next is, in conjunction with a number of critical biologically-motivated network properties and mechanisms, capable of generating abstract, invariant categorical representations of the overall shapes of objects, with a structure that matches human shape similarity judgments. Thus, we have demonstrated that predictive learning has the potential to go beyond the surface structure of its inputs, and develop systematic, abstract encodings of the deeper structure of the environment. Considerable further work remains to be done to more precisely characterize the essential properties of our biologically-motivated model necessary to produce this abstract form of learning, and to further explore the full scope of abstract learning across different domains. We strongly suspect that extensive cross-modal predictive learning in real-world environments is a significant factor in human and specifically infant development and could greatly multiply the opportunities for the formation of higher-order abstract representations that more compactly and systematically capture the structure of the world. Furthermore, extending predictive learning to encompass motor interactions with the environment can potentially create much richer opportunities for deeper learning. Future versions of these models could thus potentially answer the question of how deep an understanding a pre-verbal human, or a non-verbal primate, can extract from its environment through such predictive learning mechanisms. This would then represent the foundation upon which language and cultural learning builds, to shape the full extent of human intelligence.

References

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Elman, J. L. (1990). Finding Structure In Time. *Cognitive Science*, *14*(2), 179–211.

Mumford, D. (1991). On the computational architecture of the neocortex. *Biological Cybernetics*, *65*(2), 135–145.

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. (1998). Six Principles for Biologically-Based Computational Models of Cortical Cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.

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.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(9), 533–536.

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

Shipp, S. (2003). The functional logic of cortico-pulvinar connections. *Philosophical Transactions of the Royal Society of London B*, *358*(1438), 1605–1624.

Ungerleider, L. G., & Mishkin, M. (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *The Analysis of Visual Behavior* (pp. 549–586). Cambridge, MA: MIT Press.