Neocortical plasticity: an unsupervised cake but no free lunch

Eilif B. Muller*

Philippe Beaudoin
Element AI
6650 Saint-Urbain #500
Montreal, QC H2S 3G9
Canada

*Correspondence to: eilif.muller@elementai.com

Abstract

The fields of artificial intelligence and neuroscience have a long history of fertile bi-directional interactions. On the one hand, important inspiration for the development of artificial intelligence systems has come from the study of natural systems of intelligence, the mammalian neocortex in particular. On the other, important inspiration for models and theories of the brain have emerged from artificial intelligence research. A central question at the intersection of these two areas is concerned with the processes by which neocortex learns, and the extent to which they are analogous to the back-propagation training algorithm of deep networks. Matching the data efficiency, transfer and generalization properties of neocortical learning remains an area of active research in the field of deep learning. Recent advances in our understanding of neuronal, synaptic and dendritic physiology of the neocortex suggest new approaches for unsupervised representation learning, perhaps through a new class of objective functions, which could act alongside or in lieu of back-propagation. Such local learning rules have implicit rather than explicit objectives with respect to the training data, facilitating domain adaptation and generalization. Incorporating them into deep networks for representation learning could better leverage unlabelled datasets to offer significant improvements in data efficiency of downstream supervised readout learning, and reduce susceptibility to adversarial perturbations, at the cost of a more restricted domain of applicability.

Unsupervised neocortex

The neocortex is the canonically 6-layered sheet of cells forming the grey matter surface of the mammalian cerebrum. It is composed of a densely interconnected network of sub-regions responsible for learning sensory processing, speech and language, motor planning and many of the higher cognitive processes associated with rational thought. The human neocortex contains an estimated 100 trillion synapses, the points of communication between neurons which undergo persistent changes in strength and topology as a function of signals local to the synapse and a complex biochemical program (Holtmaat and Svoboda, 2009). These processes, broadly known as synaptic plasticity, are thought to be the basis of learning and memory in the brain.

An important task of synaptic plasticity in sensory neocortical areas is to learn disentangled invariant representations (DiCarlo et al., 2012). For example, the ventral stream of primate visual cortex,

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

the collection of areas responsible for visual object recognition, computes hierarchically organized representations much like state-of-the art convolutional neural networks (CNNs) optimized for the task (Yamins et al., 2014).

While there are impressive similarities in the learned representations between the ventral stream and CNNs, there are important differences in *how* those representations are learned. While CNNs are trained in a supervised manner using a gradient descent optimization algorithm with an explicit global objective on large labelled datasets, the ventral stream learns from a much larger dataset (visual experience) but with only very sparse labelling. The latter property of cortical learning is attractive to emulate in CNNs, and more broadly across deep learning models. Attractive, not only because of the ability to make use of unlabelled data during learning, but also because it will impart the models with superior generalization and transfer properties, as discussed below.

The monkey's paw effect: the problem with specifying what without specifying how

A well known and often encountered pitfall of numerical optimization algorithms for high dimensional problems, such as evolutionary algorithms, simulated annealing and also gradient descent, is that they regularly yield solutions matching *what* your objective specifies to the letter, but far from *how* you intended (Lehman et al., 2018).

The short story "The Monkey's Paw" by W. W. Jacobs provides a compelling metaphor. In that story, the new owner of a magical mummified monkey's paw of Indian origin is granted three wishes. The owner first wishes for $200, and his wish is eventually granted to the penny, but with the grave side effect that it is granted through a goodwill payment from his son's employer in response to his untimely death in a terrible machinery accident (Jacobs and Parker, 1910).

The Monkey's Paw effect is also applicable to gradient descent-based optimization of deep neural nets. The relative data-hungriness of current supervised learning strategies, and the use of data augmentation to improve generalization reflect the precarious position we are in of needing to micromanage the learning processes.

Adversarial examples (Moosavi-Dezfooli et al., 2016) are evidence that the monkey's paw effect none-the-less persists. It is temping to continue with the current paradigm and re-inject adversarial examples back into the learning data stream. Extrapolating, this goes in the direction of specifying the negative space of the objective, all those things the optimization should not do to solve the problem, which is potentially infinite, and rather risky in production environments like self-driving cars.

Adversarial examples represent an opportunity to address the issue in a more fundamental way (Yamins and DiCarlo, 2016). It has been argued by Bengio (2012) that if we could design deep learning systems with the explicit objective of "disentangling the underlying factors of variation" in an unsupervised manner, then there is much to be gained for generalization and transfer.

Such an approach offers a promising solution to the Monkey's Paw effect, as there is an explicit objective of learning good representations, from which generalization and transfer follow by definition.[1] One small challenge remains: how to express the objective of learning good representations? If we restrict ourselves to the subset of all possible inputs for which the neocortex learns good representations, the local processes of synaptic plasticity may provide valuable clues.

Neocortical plasticity

The neocognitron model (Fukushima, 1980), the original CNN architecture, learned visual features through self-organization using local rules. Since its conception, our understanding of the neocortex and its neurons and synapses has progressed considerably.

Recent insights into the local plasticity rules for learning in the neocortex offer new inspiration for deep representation learning paradigms that learn "disentangled representations" from large unlabelled datasets in an unsupervised manner. A selection of recent insights into the systems of plasticity of the neocortex is shown in Fig. 1. A new dendrite-centric view of synaptic plasticity is

[1] For some input spaces, such as white noise, a good representation may be undefined.

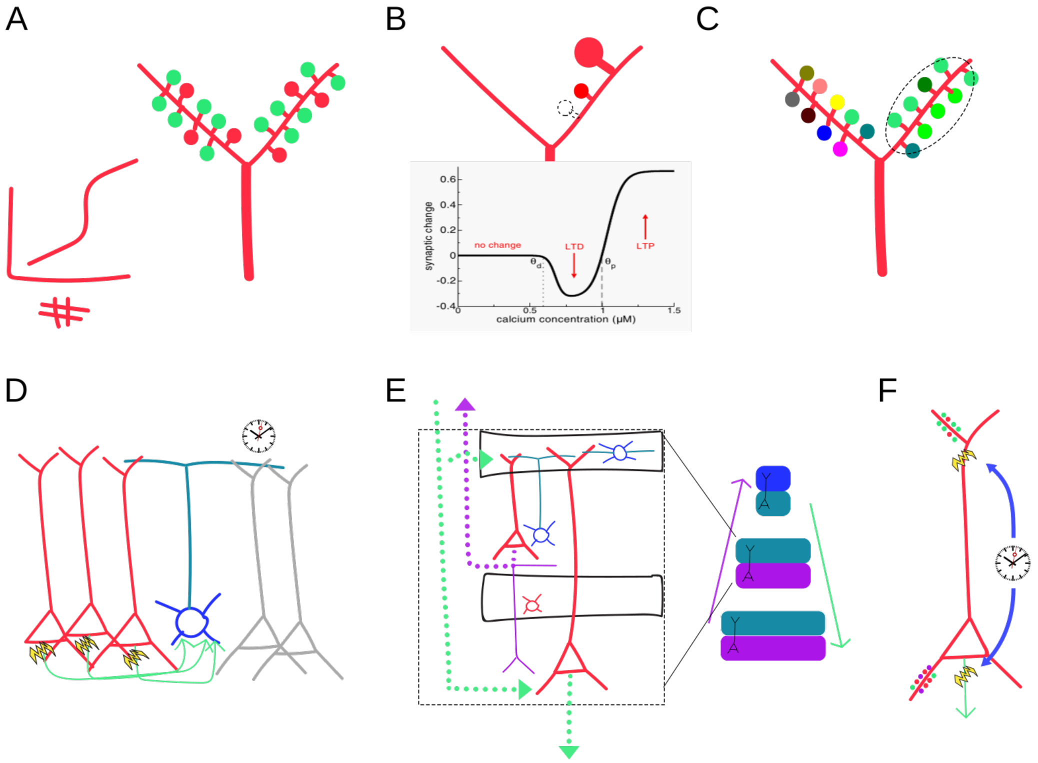

Figure 1: **A selection of recent insights into the dendritic mechanisms of plasticity of the neocortex.** (**A**) Concurrent activation of > 10 nearby synapses in pyramidal neuron dendrites (red) triggers NMDA plateau potentials in dendrites (left). (**B**) Calcium drives synaptic plasticity. Synapses are bi-stable, and can be added or removed in the weak state (above). NMDA plateau potentials drive potentiation of synapses through their associated large calcium currents. (source: Graupner and Brunel (2010)) (**C**): Clusters of co-coding synapses are captured through these mechanisms. (**D**) Co-coding neurons form small cliques, reinforced through cluster capture. These cliques activate Martinotti cells which block further capture, implementing opposing competition. (**E**) Neocortical areas are organized in a hierarchy with top-down input arriving in layer 1 (the top-most layer) at the apical tufts of pyramidal dendrites, and at layer 6 and lower layer 5. (**F**). Temporal association of top-down and bottom-up drives cliques and plasticity.

emerging with the discovery of the NMDA spike, a non-linear mechanism hypothesized to associate co-activated synapses through potentiation or structural changes driven by the resulting calcium currents (Schiller et al., 2000; Graupner and Brunel, 2010; Holtmaat and Svoboda, 2009) (Fig. 1A-B). Such associations, in the form of co-coding clusters of synapses, have recently been experimentally observed using optical techniques (Wilson et al., 2016) (Fig. 1C). Moreover neurons in the neocortex are known to form small cliques of all-to-all connected neurons which drive co-coding (Reimann et al., 2017), a process that would be self-reinforced through dendritic clustering by NMDA spikes (Fig. 1D). Martinotti neurons, which are activated by such cliques of pyramidal neurons, and subsequently inhibit pyramidal dendrites (Silberberg and Markram, 2007) provide well-timed inhibition to block further NMDA spikes (Doron et al., 2017), and put a limit on the maximal pyramidal clique size, but also suppress activation of competing cliques (e.g. Winner-take-all (WTA) dynamics). Together, such plasticity mechanisms appear to form basic building blocks for representation learning in the feedforward pathway of the neocortex using local learning rules. While long known competitive strategies for unsupervised representation learning indeed rely on WTA dynamics (Fukushima, 1980; Rumelhart and Zipser, 1985), deep learning approaches incorporating these increasingly apparent dendritic dimensions of learning processes have yet to be proposed (Poirazi and Mel, 2001; Kastellakis et al., 2015).

Unlike CNNs, the neocortex also has a prominent feedback pathway down the hierarchy, whereby top-down input from upper layers innervate the apical tufts of pyramidal cells in layer 1 of a given cortical region (Felleman and Van, 1991). Associations between top-down and feed-forward (bottom-up) activation are known to trigger dendritic calcium spikes and dendritic bursting (Larkum et al., 1999), which again specifically activates the WTA dynamics of the Martinotti neurons (Murayama et al., 2009), but disinhibitory VIP neurons can also modulate their impact (Karnani et al., 2016). These feed-back pathways have been proposed to implement *predictive coding* (Rao and Ballard, 1999), and error back-propagation for supervised learning algorithms (Guerguiev et al., 2017; Sacramento et al., 2018). While their importance for rapid object recognition has been recently demonstrated, their computational role remained inconclusive (Kar et al., 2019).

Cake but no free lunch

With the demonstrated applicability of supervised learning for a broad range of problems and data distributions, and an ever expanding toolbox of optimized software libraries, it is unlikely that supervised learning, back-propagation and gradient descent will be dethroned as the work horses of AI for many years to come.

Nonetheless, as applications of deep networks are moving into regions where sparse data, generalization and transfer are increasingly important, unsupervised approaches designed with the explicit goal of learning good representations from mere observation may find an important place in the AI ecosystem.

Quoting Yann LeCun[2]

> “If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning, and the cherry on the cake is reinforcement learning.”

A promising strategy would be to assume learning with sparse labels, overcoming adversarial examples, transfer learning, and few-shot learning together as the success criteria for the further development of the powerful unsupervised approaches we seek.

Recent advances in our understanding of the processes of neocortical plasticity may well offer useful inspiration, but let’s close with some words of moderation. Biology’s solutions also show us there will be no free lunch, i.e. neocortical unsupervised learning algorithms will be less general than supervised learning by gradient descent. Neocortex relies on structure at specific spatial and temporal scales in its input streams to learn representations. Evolution has had millions of years to configure the sensory organs to provide signals to the neocortex in ways that it can make sense of them, and that serve the animal’s ecological niche. We should not expect, for example, cortical unsupervised learning algorithms to cluster frozen white noise images. A neocortical solution requires a neocortical problem (e.g. from the so-called “Brain set” (Richards et al., 2019)), so if we are to successfully take inspiration from it, we must also work within its limitations.

Acknowledgments

Thanks to Giuseppe Chindemi, Perouz Taslakian, Pau Rodriguez, Isabeau Prémont-Schwarz, Hector Palacios Verdes, Pierre-André Noël, Nicolas Chapados, Blake Richards, and Guillaume Lajoie for helpful discussions.

References

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

Doron, M., Chindemi, G., Muller, E., Markram, H., and Segev, I. (2017). Timed synaptic inhibition shapes nmda spikes, influencing local dendritic processing and global i/o properties of cortical neurons. *Cell reports*, 21(6):1550–1561.

[2] `https://medium.com/syncedreview/yann-lecun-cake-analogy-2-0-a361da560dae`

Felleman, D. J. and Van, D. E. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

Graupner, M. and Brunel, N. (2010). Mechanisms of induction and maintenance of spike-timing dependent plasticity in biophysical synapse models. *Frontiers in computational neuroscience*, 4:136.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *ELife*, 6:e22901.

Holtmaat, A. and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, 10(9):647.

Jacobs, W. W. and Parker, L. N. (1910). *The monkey's paw: A Story in Three Scenes*. Samuel French, Inc.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974.

Karnani, M. M., Jackson, J., Ayzenshtat, I., Tucciarone, J., Manoocheri, K., Snider, W. G., and Yuste, R. (2016). Cooperative subnetworks of molecularly similar interneurons in mouse neocortex. *Neuron*, 90(1):86–100.

Kastellakis, G., Cai, D. J., Mednick, S. C., Silva, A. J., and Poirazi, P. (2015). Synaptic clustering within dendrites: an emerging theory of memory formation. *Progress in neurobiology*, 126:19–35.

Larkum, M. E., Zhu, J. J., and Sakmann, B. (1999). A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, 398(6725):338.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., et al. (2018). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv preprint arXiv:1803.03453*.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Murayama, M., Pérez-Garci, E., Nevian, T., Bock, T., Senn, W., and Larkum, M. E. (2009). Dendritic encoding of sensory stimuli controlled by deep cortical interneurons. *Nature*, 457(7233):1137.

Poirazi, P. and Mel, B. W. (2001). Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron*, 29(3):779–796.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79.

Reimann, M. W., Nolte, M., Scolamiero, M., Turner, K., Perin, R., Chindemi, G., Dłotko, P., Levi, R., Hess, K., and Markram, H. (2017). Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in computational neuroscience*, 11:48.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Ponte Costa, R., De Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22:1761–1770.

Rumelhart, D. E. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive science*, 9(1):75–112.

Sacramento, J., Costa, R. P., Bengio, Y., and Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in Neural Information Processing Systems*, pages 8721–8732.

Schiller, J., Major, G., Koester, H. J., and Schiller, Y. (2000). Nmda spikes in basal dendrites of cortical pyramidal neurons. *Nature*, 404(6775):285.

Silberberg, G. and Markram, H. (2007). Disynaptic inhibition between neocortical pyramidal cells mediated by martinotti cells. *Neuron*, 53(5):735–746.

Wilson, D. E., Whitney, D. E., Scholl, B., and Fitzpatrick, D. (2016). Orientation selectivity and the functional clustering of synaptic inputs in primary visual cortex. *Nature neuroscience*, 19(8):1003.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.