Computational Neuroscience Offers Hints for More General Machine Learning

David Rawlinson(✉) and Gideon Kowadlo

Incubator 491 Pty Ltd, Melbourne, Australia
{dave,gideon}@agi.io

Abstract. Machine Learning has traditionally focused on narrow artificial intelligence - solutions for specific problems. Despite this, we observe two trends in the state-of-the-art: One, increasing architectural homogeneity in algorithms and models. Two, algorithms having more general application: New techniques often beat many benchmarks simultaneously. We review the changes responsible for these trends and look to computational neuroscience literature to anticipate future progress.

Keywords: Machine learning · Biological plausibility · Credit assignment problem · Reinforcement learning · Spike timing dependent plasticity · Sparse coding · Predictive coding

1 Introduction

While Machine Learning research has traditionally focused on Narrow AI tasks, state-of-the-art solutions have become more homogeneous and generally applicable. This paper will review these trends and look to computational neuroscience for tips on future changes.

Contrast object recognition in 2005 with today. We have moved from designer architectures of specialized components to homogeneous deep networks. The old way to recognize objects was to combine explicit feature detectors such as HoG [1] with techniques like RANSAC [2] to find concensus about their geometric relationships. The new way is simply to expose a homogeneous deep, convolutional, region proposal network [3] to a very large set of labelled training images to segment objects from background.

2 Biological Plausibility of Artificial Neural Networks

The most biologically-implausible features of current supervised artificial neural networks are also related to some of their practical limitations.

The Credit Assignment Problem concerns synchronized back-propagation of error gradients from the output to hidden layer cells that caused them [4]. Although feedback connections outnumber feedforward in cortical neural networks [5,6], a biological basis for deep backpropagation is unlikely [7] because

T. Everitt et al. (Eds.): AGI 2017, LNAI 10414, pp. 123–132, 2017.
DOI: 10.1007/978-3-319-63703-7_12

it requires dense and precise reciprocal connections between neurons. Biological feedback connections also modulate or drive output, whereas in feed-forward artificial networks, feedback is only used for training. Layerwise backpropagation of error gradients is not supported by biological evidence [8]. Credit assignment also poses practical computational problems such as vanishing and exploding (shattered) gradients [9], which can limit network depth. Since current theory suggests that deeper is better than wider [10], this is a major problem. Recent work on decoupled neural interfaces looks to avoid this limitation via local cost functions [11].

Credit assignment is also difficult when inputs and outputs are separated by time. Only recurrent networks with gated memory cells as used in LSTM [12] have enabled effective back-propagation of error gradients over longer periods of time. There is some evidence for a biological equivalent of memory and gates in biological neurons [13].

Modern artificial neural networks appear to have enough capacity, but improved generalization seems to require better regularization of network weights [14]. Currently this tends to be explicit, e.g. weight-decay, but it could be implicit in better models.

Supervised training requires large, labelled training datasets. For embodied agents this is problematic: it is necessary to generate correct output for the agent in all circumstances. This appears to be at least as difficult as building the agent control system. Reinforcement learning avoids this problem by providing feedback about an agents output or actions in via an abstract "reward" signal. The ideal output is not required. One of the most popular reinforcement learning methods is Q-learning [15]. $Q(s, a)$ is defined as the maximum discounted future reward of performing action a in state s.

The task of associating current actions with future rewards is normally tackled via discounting. There is considerable biological evidence to support the hyperbolic temporal discounting model for associating causes with outcomes, including fMRI studies [16], recordings of neuron activity [17], and behavioural studies [18] (including human) [19].

But as always, there are practical problems. Although Q-learning is guaranteed to converge on true Q-values given training samples in any order, we cannot know how close we are while some actions are unexplored. We need a policy to balance exploration (discovery of accurate Q values) versus exploitation of current Q values. We can find some heuristic guidance from e.g. animal studies of foraging exploration behaviour [19], but in a naive representation the space of all possible states and actions is simply too large to be practically explored. As representations become more sophisticated, the gaps between sampled rewards become larger. We need a way to generalize from a smaller set of experiences.

There are two approaches to this generalization problem [20]. First, we can try to generalize explicitly by predicting commonalities between states and then inferring rewards. Second, we can try to reduce the dimensionality of the state-space by creating more abstract, hierarchical representations of the data.

Impressive results were achieved by Mnih et al. [21] playing Atari games with their Deep Q Learning method. They used several tricks to overcome exploration and representation limits. First, they built a smaller, hierarchical state-space in which it is easier to learn Q values. Second, they used "experience replay" to accelerate Q value training. Third, they used ϵ -greedy exploration to balance exploration and exploitation.

But the key to improving the generalization of observed discounted rewards may lie in other aspects of human general intelligence, such as attentional strategies. Using working memory [22] and attentional gating, the current state can become a filtered construction of features relevant to the problem under consideration, even if they were not observed in the immediate past (working memory allows humans to store several items for a few minutes at most).

Graves et al. recently published the "Neural Turing Machine", combining recurrent neural networks with a memory system [23]. The architecture is described in a very mechanical way, with a tape-like memory store and read/write heads - hence the name, which is derived from the original Turing machine concept. But despite the mechanical description, their intention was to simulate the properties of human working memory in a differentiable architecture that could be trained by gradient descent. They demonstrated that the system could perform a number of computational tasks, even a priority-sort. Zaremba and Sutskever later extended the concept to reinforcement learning and reproduced some of the tasks [24].

Overall, how closely do artificial neural networks match the computational properties of natural ones? There is neurological evidence of computational similarities between machine learning and human general intelligence [25]. Similar visual feature detectors can be learned [26], and the same types of variation are confusing for both deep learning and people [27]. But the discovery of adversarial examples [28], which look ordinary to us but confusing to artificial networks, suggests a weakness in artificial representations. Surprisingly, the deficiency seems to be fundamental to models produced by training linear discriminators in high dimensional spaces [29], because the problem "generalizes" to unseen data, and disjoint training instances are vulnerable to the same perturbations!

3 Interesting Features of Biological Neural Networks

Biological neurons are more complex than conventional artificial ones and are believed to learn using different, mostly local rules, such as Spike-Timing Dependent Plasticity (STDP) [30], pre & post synaptic correlation [31] inter-cellular competition [32] and lifetime sparsity (firing rate) constraints [33, 34], as opposed to global error minimization. Recently, neuroscientists have become interested in the role of dendrite computation: Individual neurons can have many layers of branching, and a transfer function between branches [35]. This means that an individual biological neuron can be computationally equivalent to a tree of conventional artificial neurons. Within a neuron, precise feedback could exist, allowing supervised training of a few layers against local cost functions.

In machine learning, recent progress has also concerned increasingly sophisticated components, such as the Inception module [36]. These may be more biologically realistic that at first it appears, although Inception does propagate gradients between modules.

Neuroscientists are convinced that the Spike-Timing Dependent Plasticity (STDP) rule accurately describes the way many neurons synapses adjust their weights [37]. This is an unsupervised rule; weights are adapted to strengthen synapses that reliably predict a post-synaptic (output) spike before it occurs. This is computationally convenient due to use of only local information during learning. Interestingly, artificial simulations of neurons including recurrent connections and STDP learning are able to produce some of the best wholly unsupervised representations. For example, Diehl & Cook [38] were able to achieve 95% classification accuracy on MNIST using unsupervised learning and cell-label correlation: The training data was used to correlate spiking of each cell with the ten training data digit labels. Each cell therefore had equal influence on the classification result.

Normally, such high-performance classification requires a supervised layer to optimize decision boundaries, allowing the influence of each feature or cell to be varied; that they were able to achieve this without this type of supervised training implies that their technique was very effective in capturing general structure of the input.

State of the art results in machine learning are nowadays typically produced by supervised learning. Yet researchers have recognized that unsupervised pre-training of shallower layers improves performance of a larger supervised network [39]. Improved theoretical understanding of this suggests that unsupervised learning acts as a regularizer, producing features that generalize better [39,40].

Historically, unsupervised layers required greedy layerwise pre-training. However, use of linear (e.g. the Rectified Linear (ReLU)) transfer function [41], and techniques like batch normalization [42] have allowed simultaneous training of all layers in deep networks. In fact, continuing the trend of increasing simplicity, Exponential Linear Units (ELUs) [43] aim to avoid the need for statistical preprocessing such as batch normalization, although not yet completely.

Biological neural networks continue to outperform artificial ones in several learning characteristics. Optimal modelling of nonstationary input [44] is particularly relevant to General Intelligence tasks featuring an embodied agent, whose choices can suddenly and permanently change input statistics. For example, if you suddenly start to explore a new part of the environment, you will see new things. A robot that leaves the lab to explore the gardens will now frequently encounter trees and plants, that are completely unlike all previous visual perceptions of right-angled corridors, doors and desks. This is problematic when training has moved weights into unsuitable ranges for further learning, but work on strategies such as adaptive learning rates will likely help [45].

Biological neural networks learn both more quickly [34] and more slowly [46] than artificial ones. Although this seems contradictory, this flexibility is actually a desirable quality. In psychology, “one trial” or “one shot” learning occurs

when just one experience is enough to modify future behaviour. In artificial intelligence, one way to tackle this is instance-based learning [47], in which all training samples are stored; the model is implicitly generated by interpolating between these samples. Unfortunately, this approach does not scale. Other prominent examples of "one shot" learning include [48,49]. Recently, Santoro et al. [50] developed a one-shot method using Memory-Augmented Neural Networks (such as the Neural Turing Machine described above). In this system, a differentiable architecture is trained to solve the meta-problem of using an external memory to store new learnings after a single presentation. Interestingly, an interaction between working memory in the ventrolateral prefrontal cortex and the hippocampus, may be used as a switch to activate one-shot learning in the brain [51].

Common training methods in machine learning such as gradient descent must learn slowly to avoid catastrophic interference between new and existing weights. Until recently, it was thought that synapse formation was also a relatively slow and permanent process. Artificial neural networks are typically modelled with fixed synaptic connectivity, only varying the synaptic efficiency, or "weight", of each connection. But more recently, it was confirmed that synapses are actually formed quickly and throughout adult life, perhaps in response to a homeostatic learning rule that controls lifetime firing rate [34].

Given all the above, even slower learning might seem undesirable; but to improve representation we need to reduce bias towards recent samples while continuously integrating new information. This relatively unexplored topic is known as lifelong machine learning [52].

4 Sparse and Predictive Coding

The encoding of electrical information transmitted between neurons is under active investigation. Neurons normally fire in bursts and trains of varying frequency, with long rests in between. The meaning of these spike trains is uncertain [53], but STDP learning rules are sensitive to spike timing and rate.

We also observe that most neurons are silent most of the time - only a fraction will fire at any given moment. This phenomenon is loosely defined in neuroscience as sparse coding [54]. Sparse coding has a number of theoretical advantages, such as combinatorial representational power and a natural, robust similarity metric in the intersection of active cells.

In machine learning sparse coding has a more specific meaning, namely the learning of a set of overcomplete basis vectors representing the input [54]. Research shows that deep unsupervised sparse coding produces very useful features: Le et al. [55] created a deep sparse architecture that spontaneously functioned as a high accuracy face detector without being optimized to perform this task.

Why does sparse coding help? It is believed that the process of sparse encoding is an inherently superior form of dimensionality reduction compared to e.g. vector quantization [56]. How you train the overcomplete bases is surprisingly

less important, assuming the input has been normalized. Sparse coding chooses a few bases that jointly describe each input combination, meaning that subsets may be used in novel combinations without retraining. This is also known as the "union property" of sparse representations [57]. Again we observe that a representational change has provided computationally beneficial effects: Which is interesting, because neuroscience also offers another representational change we might adopt - predictive coding [58].

Predictive coding proposes that cortical cells internally predict either their input or activity within local populations, and then emit signals representing prediction errors [59]. This changes the inter-layer signal from a representation of the input, to a representation of cells' inability to predict the input. Internally, only the relationships between errors are represented. This is very efficient: The representation adapts to the characteristics of the input. Input that can't be "explained" (i.e. predicted) is relayed to other layers for processing, in hope that additional resources or data will help. Errors propagate towards features that can explain them. Note the assumption here, that being able to predict an observation means that its causal relationships are being modelled correctly and thus is it "explained".

This characteristic is reminiscent of Highway Networks [60] and Deep Residual Learning (DRL) [61]. In DRL, the output of a bi-layer module is added to the module's input, requiring the module to learn any residual error between the input and desired output. DRL propagates residual error gradients in the feedback direction during training. Highway Networks, derived from LSTM, use an explicit gating mechanism to determine propagation of the input.

In both Highway Networks and DRL, modules are trained to decide to what extent they involve themselves in current input. Signals can be relayed or modified depending on the capabilities and relevance of the local module.

The reason DRL works lies in the details of the training problem posed by this architecture. Each module's output contributes additively rather than multiplicatively, and in consequence data flows freely into very deep hierarchies. After training, DRL networks become ensembles of shallower networks [62]. Just as in predictive coding, input data dynamically determines the effective depth of the network.

5 Conclusion

We have observed increasing homogeneity and generality in state-of-the-art machine learning. Biology continues to inspire this process, such as Liao and Poggio's combination of deep residual and recurrent networks [63]. In future we expect to see increasing use of local [11] and unsupervised learning rules [8], modularized architectures that promote data-driven deep representations [60,61] and dramatic improvements in the representation of state and action spaces in reinforcement learning to overcome the generalization problem.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
2. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 91–99 (2016)
4. Luo, H., Fu, J., Glass, J.: Bidirectional backpropagation: towards biologically plausible error signal transmission in neural networks. arXiv preprint arXiv:1702.07097 (2017)
5. Petro, L.S., Vizioli, L., Muckli, L.: Contributions of cortical feedback to sensory processing in primary visual cortex. Front. Psychol. **5**, 1–8 (2014)
6. 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.: Anatomy of hierarchy: feedforward and feedback pathways in Macaque visual cortex. J. Comp. Neurol. **522**, 225–259 (2014)
7. Balduzzi, D., Vanchinathan, H., Buhmann, J.: Kickback cuts Backprops red-tape: biologically plausible credit assignment in neural networks. In: 9th AAAI Conference on Artificial Intelligence (2014)
8. Bengio, Y., Lee, D.H., Bornschein, J., Mesnard, T., Lin, Z.: Towards biologically plausible deep learning. arXiv preprint arXiv:1502.04156v3 (2016)
9. Balduzzi, D., Frean, M., Leary, L., Lewis, J.P., Ma, K.W.D., McWilliams, B.: The shattered gradients problem: if resnets are the answer, then what is the question? arXiv preprint arXiv:1702.08591 (2017)
10. Eldan, R., Shamir, O.: The power of depth for feedforward neural networks. In: 29th Annual Conference on Learning Theory, PMLR, vol. 49, pp. 907–940 (2016)
11. Jaderberg, M., Czarnecki, W.M., Osindero, S., Vinyals, O., Graves, A., Kavukcuoglu, K.: Decoupled neural interfaces using synthetic gradients. arXiv preprint arXiv:1608.05343 (2016)
12. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
13. Monner, D.D., Reggia, J.A.: Systematically grounding language through vision in a neural network. In: Artificial General Intelligence: 4th International Conference (2011)
14. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: International Conference on Learning Representations (2017)
15. Kaelbling, L.P., Moore, A.W.: Reinforcement learning: a survey. J. Artif. Intell. Res. **4**, 237–285 (1996)
16. McClure, S.M., Laibson, D.I., Loewenstein, G., Cohen, J.D.: Separate neural systems value immediate and delayed monetary rewards. Science **306**(5695), 503–507 (2004)
17. Kim, S., Hwang, J., Seo, H., Lee, D.: Valuation of uncertain and delayed rewards in primate prefrontal cortex. Neural Netw. **22**(3), 294–304 (2009)
18. Hwang, J., Kim, S., Lee, D.: Temporal discounting and inter-temporal choice in rhesus monkeys. Front. Behav. Neurosci. **3**(9), 1–13 (2009)

19. Namboodiria, V.M.K., Levyc, J.M., Mihalasd, S., Simse, D.W., Shulerc, M.G.H.: Rationalizing spatial exploration patterns of wild animals and humans through a temporal discounting framework. In: Proceedings of the National Academy of Sciences, vol. 113, no. 31, pp. 8747–8752 (2016)
20. Ponsen, M., Taylor, M.E., Tuyls, K.: Abstraction and generalization in reinforcement learning: a summary and framework. In: Taylor, M.E., Tuyls, K. (eds.) ALA 2009. LNCS, vol. 5924, pp. 1–32. Springer, Heidelberg (2010). doi:10.1007/978-3-642-11814-2_1
21. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)
22. Baddeley, A., Eysenck, M., Anderson, M.: Memory, chap. 3. Psychology Press, New York (2009)
23. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401v2 (2014)
24. Zaremba, W., Sutskever, I.: Reinforcement learning neural turing machines. In: Proceedings of ICLR 2016 (2016)
25. Marblestone, A., Wayne, G., Kording, K.: Toward an integration of deep learning and neuroscience. Front. Comput. Neurosci. **7**, 137 (2016)
26. Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M., Masquelier, T.: Deep networks can resemble human feed-forward vision in invariant object recognition. Nat. Sci. R **6** (2016). Article 32672
27. Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M., Masquelier, T.: Humans and deep networks largely agree on which kinds of variation make object recognition harder. Front. Comput. Neurosci. **10**, 92 (2016)
28. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)
29. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
30. Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., Wu, Y.: STDP as presynaptic activity times rate of change of postsynaptic activity. arXiv preprint arXiv:1509.05936 (2015)
31. Oja, E.: A simplified neuron model as a principal component analyzer. J. Math. Biol. **15**, 267–273 (1982)
32. Rumelhart, D. E., McClelland, J. L., The PDP research group: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1 and 2. MIT Press, Cambridge (1986)
33. Makhzani, A., Frey, B.: Winner-take-all autoencoders. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), pp. 2791–2799 (2015)
34. Butza, M.: Brain Research Reviews **60**(2), 287–305 (2009)
35. Guergiuev, J., Lillicrap, T.P., Richards, B.A.: Towards deep learning with segregated dendrites. arXiv preprint arXiv:1610.00161 (2016)
36. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
37. Sjstrm, P.J., Rancz, E.A., Roth, A., Husser, M.: Excitability, dendritic, plasticity, synaptic. Physiol. Rev. **88**, 769–840 (2008)
38. Diehl, P., Cook, M.: Unsupervised learning of digit recognition using spike-timing-dependent plasticity. Front. Comput. Neurosci. **9** (2015). Article 99

39. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P.: Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. **11**, 625–660 (2010)
40. Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., Vincent, P.: The difficulty of training deep architectures and the effect of unsupervised pre-training. In: International Conference on Artificial Intelligence and Statistics, vol. 5, pp. 153–160 (2009)
41. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS) (2012)
42. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML, vol. 1, no. 4 (2015)
43. Shah, A., Kadam, E., Shah, H., Shinde, S., Shingade, S.: Deep residual networks with exponential linear unit. arXiv preprint arXiv:1604.04112 (2016)
44. Ditzler, G., Roveri, M., Alippi, C., Polikar, R.: Learning in nonstationary environments: a survey. IEEE Comput. Intell. Mag. **10**(4), 12–25 (2015)
45. Schaul, T., Zhang, S., LeCun, Y.: No more Pesky learning rates. In: Proceedings of 30th International Conference on Machine Learning (ICML) (2013)
46. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumarana, D., Hadsella, R.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. **114**(13), 3521–3526 (2017)
47. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn, p. 733. Prentice Hall, Englewood Cliffs (2003). ISBN 0-13-080302-2
48. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI), vol. 28, no. 4, pp. 594–611 (2006)
49. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (2000)
50. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: One-shot learning with memory-augmented neural networks. arXiv preprint arXiv:1605.06065v1 (2016)
51. Lee, S.W., ODoherty, J.P., Shimojo, S.: Neural computations mediating one-shot learning in the human brain. PLoS Biol **13**(4), e1002137 (2015)
52. Silver, D., Yang, Q., Li, L.: Lifelong machine learning systems: beyond learning algorithms. In: Papers from the 2013 AAAI Spring Symposium (2013)
53. Krahe, R., Gabbiani, F.: Burst firing in sensory systems. Nat. Rev. Neurosci. **5**, 13–23 (2004)
54. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: a strategy employed by VI? Vis. Res. **37**(23), 3311–3326 (1997)
55. Le, Q.V., Ranzato, M.A., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., Ng, A.Y.: Building high-level features using large scale unsupervised learning. In: Proceedings of the 29th International Conference on Machine Learning (2012)
56. Coates, A., Ng, A.Y.: The importance of encoding versus training with sparse coding and vector quantization. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA (2011)
57. Hawkins, J., Ahmad, S.: Why neurons have thousands of synapses, a theory of sequence memory in neocortex. Front. Neural Circ. **10**(23), 1–13 (2016)
58. Kok, P., de Lange, F.P.: Predictive coding in sensory cortex. In: Forstmann, B.U., Wagenmakers, E.-J. (eds.) An Introduction to Model-Based Cognitive Neuroscience. LLC, chap. 11, vol. 221, pp. 221–244. Springer, New York (2015)

59. Kogo, N., Trengrove, C.: Is predictive coding theory articulated enough to be testable? Front. Comput. Neurosci. **9**(111), 357–381 (2015)
60. Srivastava, N., Greff, K., Schmidhuber, J.: Highway networks. In: Deep Learning Workshop (ICML 2015) (2015)
61. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
62. Veit, A., Wilber, M., Belongie, S.: Residual networks behave like ensembles of relatively shallow networks. arXiv preprint arXiv:1605.06431 (2016)
63. Liao, Q., Poggio, T.: Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. arXiv preprint arXiv:1604.03640 (2016)