Emergence of complementary learning systems through meta-learning

Zhenglong Zhou (zzhou34@sas.upenn.edu)
Anna C. Schapiro (aschapir@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA 19104, USA

Abstract:

To process information from the external world, the brain relies on a hierarchy of processing systems, which initiate in early sensory neocortical areas and converge on the hippocampus. Components of this hierarchy exhibit markedly different computational properties, with the hippocampus supporting faster plasticity and employing sparser representations. There has been extensive work on the properties of these systems, but it remains unclear how and why these systems emerged in the first place. We explore the emergence of a hierarchy of processing systems in artificial neural networks using a meta-learning approach. As networks optimize for a set of tasks, they concurrently meta-learn hyperparameters that modulate layer-wise learning rates and sparsity. We find that this meta-learning promotes superior performance, at overall higher sparsity levels. We demonstrate that key aspects of complementary learning systems emerge in the networks, with a brain-like differentiation of sparsity and learning rates across layers. Furthermore, when endowed with two pathways and trained on a task with opposing demands of individual item recognition and categorization, the models capture divergent properties between intra-hippocampal pathways. Together, these results suggest that the organization of heterogenous learning systems in the brain may arise from optimizing biological variables that govern learning rate and sparsity.

Keywords: hippocampus; neocortex; meta-learning.

Introduction

A hierarchy of sensory processing flows from neocortical areas of the brain to the hippocampus Felleman & Van Essen 1991). The hippocampus and neocortex have been posited to perform complementary computations for learning and memory: the hippocampus rapidly forms sparse, pattern-separated representations of individual experiences, while the neocortex slowly forms overlapping, distributed representations across experiences on a more extended timescale (McClelland et al., 1995). A wide range of empirical observations support these ideas (O'Reilly & Norman 2002). One possibility is that the hippocampus and neocortex are components of broader hierarchy of plasticity and sparsity (McClelland et al., 1995; Kent et al., 2016).

There has been extensive theoretical work that seeks to capture the properties of the two systems in computational models (e.g., Sun et al., 2023; Spens & Burgess 2024), which has involved manually setting up two components with different sets of assumptions; in other words, directly building the distinctions of the two systems into the models. These theories thus do not provide an account of how and why the brain arrived at its organization of distinct subsystems in the first place.

Here, we explore the emergence of a hierarchy of subsystems in artificial neural networks (ANNs). We take a meta-learning approach that builds on prior work (Gupta et al., 2020). In this approach, as ANNs optimize weights for tasks, they concurrently meta-learn hyperparameters that modulate sparsity, through a within-layer competition mechanism, and layer-specific learning rates. We observe that this approach 1) enhances computational efficiency, 2) gives rise to a brain-like hierarchical differentiation of sparsity and learning rates, and 3) enables a two-pathway model to develop divergent properties that mirror differences between intra-hippocampal pathways. Together, these results suggest that the organization of complementary learning systems in the brain may arise from meta-learning biological variables that modulate activity and plasticity.

Results

In this meta-learning approach (Gupta et al., 2020), as ANNs learn a series of tasks, they jointly optimize hyperparameters and weights through a two-level optimization process (Fig. 1). In the inner loop, the networks obtain a set of temporary “fast” weights by updating weights based on a batch of current task data. The outer loop then computes a meta loss by evaluating fast weights on a mixture of current and prior data, computes a gradient based on this loss, and uses that gradient to update initial weights (weights before inner loop updates) and hyperparameters through meta update. Our models adapt two sets of layer-wise hyperparameters. The first set consists of layer-specific learning rates. The second set modulates layer-wise sparsity through a within-layer competition mechanism (Bricken et al., 2023): during inference, the activity of the k^{th} most active unit in each layer is multiplied by a layer-specific hyperparameter before being subtracted from the activity of all units in that layer.

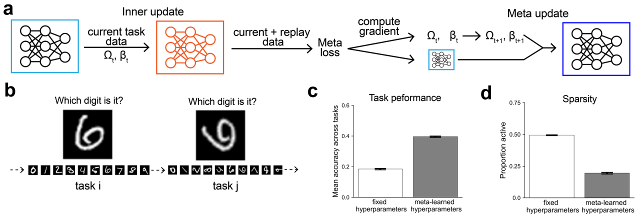

Figure 1. **Meta-learning approach and facilitation of computational efficiency**. (**a**) Schematic illustration of the two-level optimization process in the meta-learning approach. Ω denotes per-layer learning rates. β represents parameters that modulate within-layer competition. Light blue, orange, and dark blue boxes respectively denote weights before inner update, fast weights obtained through inner update, and weights after meta update. (**b**) The model learns a sequence of 20 rotated MNIST tasks, each of which rotates the standard MNIST digit dataset by a certain degree. (**c**) We first measure task performance and sparsity in feedforward networks with two hidden layers. Each hidden layer has 250 units, uses the ReLU activation function, and includes no bias term. Other simulations in this work employ the same specifications for the models but use varying numbers of layers. Relative to baseline models with matched initialization of weights and hyperparameters, models that concurrently meta-learn hyperparameters through the tasks achieve superior performance. (**d**) Compared to matched baseline models, models that meta-learn hyperparameters employ lower proportions of active units (i.e., sparser representations) to represent task inputs. Error bars represent +/-1 SEM across networks initialized with 40 random seeds.

We observed that meta-learning hyperparameters facilitates computational efficiency of feedforward neural networks: Compared to models that do not adapt hyperparameters, models that meta-learn hyperparameters achieved superior performance (Fig. 1c) while activating fewer hidden units (Fig. 1d). Unlike approaches that explicitly optimize for sparsity (Hoefler et al., 2021), this approach promotes sparse representations without directly optimizing for an energy cost or sparsity objective.

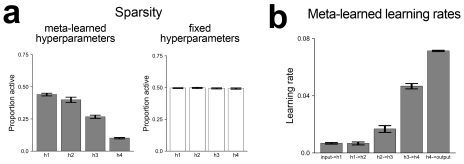

Figure 2. **A brain-like hierarchical differentiation of sparsity and learning rates emerges through meta-learning**. (**a**) In feedforward networks with 4 hidden layers that meta-learn hyperparameters, deeper layers, and especially the top hidden layer (h4), learn to employ sparser representations (left). This hierarchical differentiation of sparsity is absent in baseline models that do not meta-learn hyperparameters (right). (**b**). In models that learn hyperparameters, the top layer develops much higher learning rates than earlier layers. We note that both patterns are present in networks with two hidden layers that meta-learn hyperparameters.

Second, through this approach, a brain-like hierarchical differentiation of hidden layers emerged (Fig. 2): ANNs learned to update their incoming connections more rapidly and formed sparser task representations (i.e., activating a lower proportion of hidden units) in higher than in lower hidden layers. The emergent graded structure of these networks resembles the hierarchical organization of processing systems in the brain, with the especially sparse and fast-learning hippocampus at the apex (McClelland et al., 1995; O'Reilly & Norman 2002). Both patterns are absent in the baseline models.

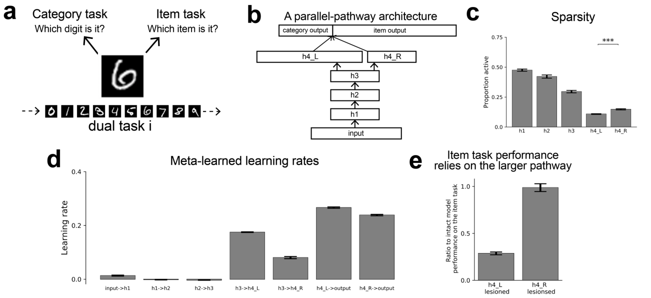

Figure 3. **The emergence of divergent properties in a model with parallel pathways and dual tasks**. (**a**) The network learns a series of dual tasks. For each dual task, the model learns to map each image to its category output (i.e., 0-9) as well as an item-specific label. (**b**) We meta-trained a model with parallel pathways of different sizes at the top of a hierarchy. h4_L (four times the size of h4_R) and h4_R (has 250 units as its preceding layers) receive input from h3. (**c**) Consistent with our previous results, the model shows a hierarchical differentiation of sparsity. In addition, the larger h4_L pathway develops sparser task representations. (**d**) The larger h4_L pathway develops higher learning rates than h4_R. (**e**) Performance on the item task relies on the fast-learning, sparse h4_L, such that lesioning h4_L dramatically impairs the item task performance. The divergent properties of the two pathways in the meta-learned model and the reliance on the larger pathway for item discrimination mirror differences between the two main intra-hippocampal pathways.

Finally, we meta-trained a model with parallel pathways of different sizes at the top of a hierarchy, representing the two main pathways (the monosynaptic and trisynaptic pathways) within the hippocampus (Fig. 3). The larger trisynaptic pathway forms sparser task representations, learns more quickly, and is essential for distinguishing exemplars (Schapiro et al., 2017; Baker et al., 2016). We trained the model on a dual task that imposes the opposing demands of categorization and distinguishing exemplars. The two pathways developed divergent properties consistent with differences between intra-hippocampal pathways, including higher learning rates and sparser task representations in the larger pathway, and a reliance on the larger pathway for distinguishing exemplars (Fig. 3).

Together, our results suggest that the organization of graded subsystems in the brain may arise from meta-learning biological variables that modulate sparsity and speed of learning. This process could potentially correspond to evolutionary, developmental, and/or concurrent optimization processes that govern online learning. The meta-learning approach provides a promising framework for understanding the organization of subsystems in the brain.

References

Baker, S., Vieweg, P., Gao, F., Gilboa, A., Wolbers, T., Black, S. E., & Rosenbaum, R. S. (2016). The human dentate gyrus plays a necessary role in discriminating new memories. Current Biology, 26(19), 2629-2634.

Bricken, T., Davies, X., Singh, D., Krotov, D., & Kreiman, G. (2023). Sparse distributed memory is a continual learner. ICLR.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991), 1(1), 1-47.

Gupta, G., Yadav, K., & Paull, L. (2020). Look-ahead meta learning for continual learning. Advances in Neural Information Processing Systems, 33, 11588-11598.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research, 22(241), 1-124.

Kent, B. A., Hvoslef-Eide, M., Saksida, L. M., & Bussey, T. J. (2016). The representational–hierarchical view of pattern separation: Not just hippocampus, not just space, not just memory?. Neurobiology of learning and memory, 129, 99-106.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological review, 102(3), 419.

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. Trends in cognitive sciences, 6(12), 505-510.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1711), 20160049.

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. Nature Human Behaviour, 1-18.

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. Nature neuroscience, 26(8), 1438-1448.