Sparsey, a memory-centric model of on-line, fixed-time, unsupervised continual learning

Rod Rinkus *
Neurithmic Systems
Newton, MA 02465, USA
rod@neurithmicsystems.com

Abstract

Four hallmarks of human intelligence are: 1) on-line, single/few-trial learning; 2) important/salient memories and knowledge are permanent over lifelong durations, though confabulation (semantically plausible retrieval errors) accrues with age; 3) the times to learn a new item and to retrieve the best-matching (most relevant) item(s) remain constant as the number of stored items grows; and 4) new items can be learned throughout life (storage capacity is never reached). No machine learning model, the vast majority of which are *optimization-centric*, i.e., learning involves optimizing a global objective (loss, energy), has all these capabilities. Here, I describe a *memory-centric* model, Sparsey, which in principle, has them all. I note prior results showing possession of Hallmarks 1 and 3 and sketch an argument, relying on hierarchy, critical periods, metaplasticity, and the recursive, compositional (part-whole) structure of natural objects/events, that it also possesses Hallmarks 2 and 4. Two of Sparsey's essential properties are: i) information is represented in the form of fixed-size sparse distributed representations (SDRs); and ii) its fixed-time learning algorithm maps more similar inputs to more highly intersecting SDRs. Thus, the similarity (statistical) structure over the inputs, not just pair-wise but in principle, of all orders present, in essence, a generative model, emerges in the pattern of intersections of the SDRs of individual inputs. Thus, semantic and episodic memory are fully superposed and semantic memory emerges as a by-product of storing episodic memories, contrasting sharply with deep learning (DL) approaches in which semantic and episodic memory are physically separate.

1 Introduction

Any human-like artificial general intelligence (AGI) must possess the hallmarks listed in the abstract. No machine learning (ML) model, including any deep learning (DL) model, has been shown to have them all. In particular, ML/DL models—the vast majority of which are *optimization-centric*, i.e., learning involves optimizing a global loss or energy objective—have been subject to *catastrophic forgetting* (CF) [24] and so have difficulty with Hallmark 2 and thus, with lifelong continual learning (CL). Equally important viz. scaling to lifelong CL, no ML/DL model has been shown to have Hallmark 3. Here, I describe a radically different AGI model for which Hallmarks 1 and 3 have already been shown, for sequences as well as purely spatial inputs. I then sketch an argument, relying on hierarchy, critical periods, metaplasticity, and the recursive, compositional (part-whole) structure of natural objects/events, that it also possesses Hallmarks 2 and 4, and further, that it can be expected to retain these properties over an effectively open-ended lifetime.

The core problem of CF is that the converged state of the weights capturing one dataset/task generally differs from that for any other. This is especially an issue for optimization-centric models, which

*www.neurithmicsystems.com and Visting Scientist, Brandeis, Biology

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

generally involve repeated, small changes along the objective's gradient for massive numbers of weights. If all weights remain permanently subject to change (across datasets/tasks), then previously stored information can be erased as the weights converge to an optimum for a new dataset/task. Several solution types have been advanced. **1.** Sparsify/orthogonalize memory traces, causing weights to be used less often, reducing competing influences on individual weights, and increasing the number of stably learnable mappings [3, 10, 38]. **2.** Continually re-present previously learned items (e.g., from old tasks) *interleaved* with new items, as in [9]. But, the set of old items needing to be interleaved continually increases, suggesting difficulty scaling to lifelong CL scenarios. However, mounting evidence (recent review in [8]) suggests the hippocampus does act as a *transient*, on-line, single-trial, learner facilitating replay of neocortical memory traces of recent experiences, allowing gradual formation of representations of higher-order statistical structure of experiences, an idea formalized in the pioneering *Complementary Learning Systems* (CLS) model [23]. **3.** Adding a regularization term to the loss function, which penalizes changing weights in proportion to their importance for previously learned mappings/tasks [1, 20]. But, this increases learning complexity since the importance measure is continually re-evaluated for every weight throughout the system's lifetime. **4.** Finally, recent approaches that add an external (episodic) memory for individual inputs to a core DL model [14, 15, 31, 40], in principle, address CF/CL in a similar vein as the CLS model: cf. [22]. However, unlike Sparsey, any model in which semantic and episodic memory are physically separate entails increased complexity in managing the interaction of the two memories including the cost of moving information between the two.

2 Brief Summary of Sparsey

Sparsey [32, 33, 35] differs fundamentally from most ML/DL models as it is not *optimization-centric*, but rather, *memory-centric*: it simply assigns/stores memory traces to inputs, e.g., successive frames of streaming video, as they present, as well as associating (chaining) those traces both sequentially in time and hierarchically (across model levels). These traces are in the form of fixed-size *sparse distributed representations* (SDRs), as in Fig. 1, which allows more similar inputs to be represented by more highly intersecting SDRs. Sparsey's learning algorithm, the *Code Selection Algorithm* (CSA) (Fig. 2, see refs for details), does *statistically* preserve similarity in this way (see Fig. 3). The similarity structure over the inputs, not just pairwise, but in principle, of all orders present, emerges automatically in the pattern of intersections over the stored SDRs. Thus, semantic memory emerges as a by-product of storing episodic memories in superposition: the same weight changes that store an episodic memory also act to create semantic memory, which as noted above, suggests a potentially large efficiency advantage over models with external episodic memories [14, 15, 31, 40]. More important, this constitutes a fundamentally different, on-line, single-trial method for building a generative model without any notion of optimization (though supervised and reinforcement learning can be easily implemented as meta-protocols).

Equally important, the CSA runs in *fixed-time*: the number of algorithmic steps needed to learn a new item remains constant as the number of stored items grows. In particular, the CSA preserves similarity *without* needing to compare new inputs to previously stored inputs, either serially, as in most nearest-neighbor models, or to a log number of them, as in tree-based models. Closest-match retrieval time is also fixed. Thus, Sparsey can be viewed as implementing, in a biologically plausible manner, similar functionality to locality-sensitive hashing (LSH) [19]. In fact, based on a recent review [39], Sparsey is more general in that: a) its simlarity metric is graded, whereas LSH's is binary; b) spatiotemporal and spatial metrics are handled qualitiatively identically; and c) the "hash index" is learned from scratch from the data, a crucial capability, especially for large datasets [21].

As Fig. 1c shows, an overall model instance is a hierarchy with each level consisting of an array of SDR coding fields (called "macs" as they are proposed analogs of the (L2/3 portions) of cortical macrocolumns), with local, bottom-up, top-down, and horizontal connectivity (most connections not shown). Three other properties (not visible in figs) are essential to the argument of Sec. 3. **1)** Event-based *critical periods*: learning in a mac (storing new SDR codes) is shut down, i.e., no further changes allowed to the mac's afferent synapses, when a threshold fraction of them have been increased. There is substantial evidence for critical periods in primary cortical areas [2, 4, 6, 7, 11, 18, 25, 27] and olfactory bulb [5, 30]. **2)** A large-delta Hebbian learning scheme combined with a *metaplasticity* concept, i.e., synaptic resistence to decay, *permanence*, denoted θ. The model starts as a *tabula rasa*: all synapses initially have w=0 and θ=0. Whenever a synapse experiences a pre-post coincidence:

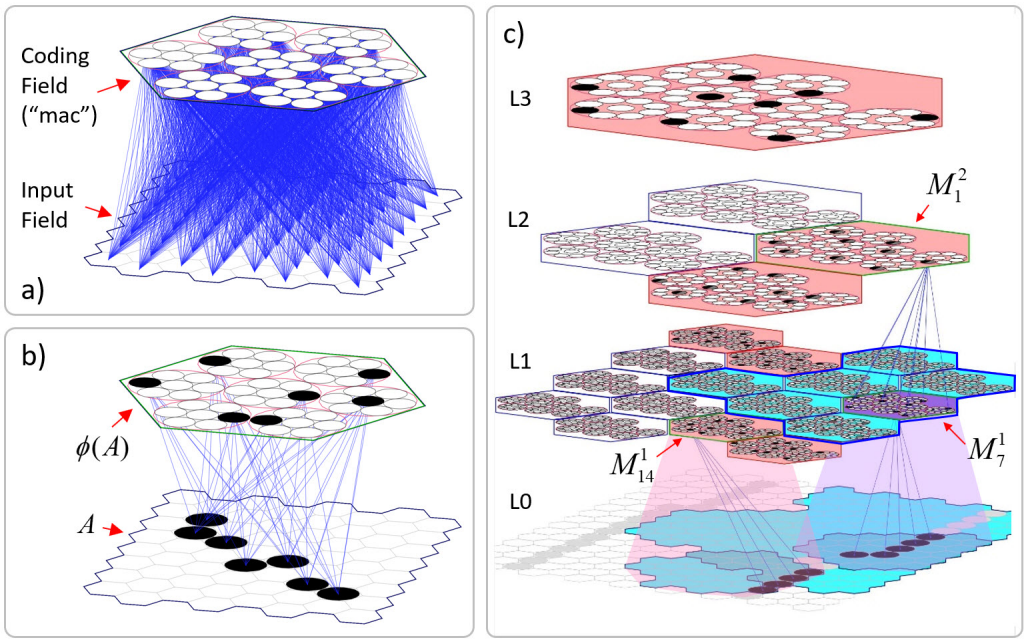

Figure 1: a) An SDR coding field ("mac") comprised of Q=7 WTA *competitive modules* (CMs), each with K=7 binary units, with full bottom-up connectivity from binary pixel input field. b) An input A, its SDR code, ϕ(A), comprised of Q active (black) units, and the weights increased (to the max, binary "1") to form the association. c) Hierarchical model showing subset of active macs (rose or violet) across all levels for single input frame of a sequence. Large cyan patch at L0 is the combined L0 receptive field (RF) of the seven L1 macs (bold blue border) comprising L2 mac M_1^2's immediate (L1) RF. Smaller L0 RFs of two L1 macs also shown (highlighted by translucent cones). Neighboring macs' RFs can overlap and the fraction of L0 "seen" by a mac increases with level due to fan-in/out.

a) its weight is (re)set to the max (binary "1"); and b) if the time since last pre-post falls within a θ-dependent time window, its θ is increased (decay rate decreased) and the window length increases [see [35] for details]. Assuming the expected time for a pre-post coincidence due to inputs reflecting a structural regularity of the world to recur is much (likely exponentially) smaller than that for a pre-post coincidence due to randomness, this metaplasticity scheme preferentially embeds SDRs (and chains of SDRs) reflecting the domain's structural regularities, i.e., its statistics. This permanence scheme is purely local, only requiring counting frames since a synapse's last weight increase, and computationally far simpler than other schemes requiring continual re-evaluation of each weight's importance to previously learned tasks/mappings, e.g., [1, 12, 20]. **3)** Unit activation duration (*persistence*) increases (e.g., doubles) with level, allowing a code at level J to associate with multiple sequentially active codes at level J-1, i.e., nested "chunking".

Figure 4 summarizes results [14] showing that an SDR-based mac has high storage capacity and can learn *complex* sequences (items can repeat many times in varying contexts, e.g., text), with single trials, addressing Hallmark 1. In recent work (unpublished), Sparsey achieved 90% accuracy on MNIST and 67% on Weizmann event recognition with unsupervised learning. The accuracy is sub-SOA, but likely much-improvable via thorough parameter space search and adding supervised learning. Nevertheless, learning is extremely fast (runs in fixed time), likely much faster than SOA methods when adjustsed for the fact that it uses no machine parallelism, thus addressing Hallmark 3.

3 Combination of Crucial Principles Yields Lifelong Continual Learning

Broadly, the 4-step argument is that the combined effects of SDR, hierarchy, imposed critical periods, metaplasticity, and the statistics of natural inputs, cause the expected time for a mac's afferent synaptic matrices to reach saturation (i.e., for the mac to reach storage capacity) to increase quickly, likely, exponentially, with the mac's hierarchical level. Thus, in practice, in a system with even few levels,

e.g., 10, as relevant to human cortex, the macs at the highest levels might never reach capacity, even over very long lifetimes operating on (or in) naturalistic domains, thus meeting Hallmark 4.

Step 1: The natural world is recursively compositional in both space in time. Objects are made of parts, which are made of sub-parts, etc. Even allowing for articulation (i.e., parts can move with respect their containing wholes), this vastly constrains the space of likely percepts compared to if all pixels of the visual field varied fully independently. Further, edges carry most of the information in images/video. Thus, Sparsey's input images (frames) are edge-filtered, binarized, and skeletonized (cheap, local operations), as in Fig. 5, further greatly constraining the space of likely percepts.

Step 2: The human visual system is hierarchical and receptive field (RF) size grows with level. Consider the 37-pixel hexagonal RFs of Fig. 6. In light of Step 1's preprocessing, Sparsey implements the policy that an L1 mac only activates if the number of active pixels in its RF is within some tight range around its RF's diameter, e.g., 6-7 pixels. This yields an input space of $C(37,6) + C(37,7) \approx 12.6M$. But, the combined effect of natural statistics and the preprocessing precludes the vast majority of those patterns, yielding a vastly smaller *likely* input space (see Figs. 6e-g).

Step 3: The highly (structurally) constrained space of inputs likely to occur in a small RF, suggests a small basis (lexicon) might plausibly represent all future inputs to the RF with sufficient fidelity to support correct inference/classification on larger-scale tasks pertinent to the overall hierarchical model. When an overall classification process is realized as a hierarchy of component classification processes occurring in many macs, whose RFs span many spatiotemporal scales, some substantial portion of the information as to the input's class resides in *which* macs are involved at each level. This decreases the accuracy needed in *individual* macs to achieve a given accuracy on the overall task, in turn, suggesting that smaller bases—entailing a greater average difference between a mac's actual input and the best-matching stored input (basis element) to which it is mapped—may suffice. Figs. 6a-d,h illustrate the basic idea. Here, assume the RF of an L2 mac (not shown) includes the whole field of depicted L1 mac RFs (green hexagons). Even with the very small basis set (Fig. 6b), a plane is clearly discernible in Fig. 6c. But, the plane is also discernible even if the active basis elements are randomly chosen (readily seen by squinting while looking at Fig. 6d), suggesting that in low-level macs, learning can be deliberately frozen (critical period terminated) even relatively early in the system's lifetime, thus preventing CF in those macs, while allowing acceptable accuracy of higher-level inference processes as well as ongoing learning at higher levels, i.e., of new compositions of frozen basis elements (features). This view is consistent with: a) emerging notions of *progressive disentangling* [13, 36, 41] in that lower-level macs will compute partial, *learned* invariances over their respective RFs, leaving progressively higher-level invariances to be handled at progressively higher levels; and b) the idea that by factoring the recognition problem into multiple scale-specific sub-problems (e.g., carried out at the different levels of a hierarchy), the number of samples needed to train each scale might be small and the number of samples needed overall might be *exponentially* smaller than for the unfactored "flat" approach, cf. viewing the ventral visual stream as reducing sample complexity [29], Bayesian belief nets [26], and Hinton's Capsules [37].

Step 4: As also suggested in Figure 1c, a level J+1 mac's RFs consists of a patch of level J macs. Once the level J macs' bases are frozen, the space of *possible* inputs to level J+1 macs is further (and, permanently) constrained. Moreover, just as an L1 mac only activates if the number of active *pixels* in its RF is within a tight range, so too, a level J+1 mac only activates if the number of *active level J macs* in its RF is within a tight range. While the space of possible raw (L0) inputs falling within the L0 RF of a level J mac increases exponentially with J, the combined effect of the above constraining *forces* allows only a tiny fraction of those possible inputs to ever occur. And, only a tiny fraction of those that do occur, occur more than once. The permanence policy acts to let memory traces of structurally produced (and thus more likely to recur) inputs become permanent, while letting traces of random/supuriopus inputs (having much longer expected times to recurrence) to fade. The effect of these principles must necessarily increase with level and my working hypothesis is that the magnitidue of the effect increases exponentially with level. That is, the rate at which inputs with sufficient novelty so as to require new memory traces (SDRs) to be assigned/stored, and thus the rate at which macs' afferent matrices approach saturation, likely decreases exponentially with level. Thus, macs at higher-level macs are *softly* protected from CF, meaning that critical periods need not be enforced at higher levels, thus allowing new, permanent learning at the higher levels for effectively open-ended lifetimes (without requiring on-line creation/allocation of new memory substrate [17, 28]), addressing Hallmark 4, and providing a solution to Grossberg's "stability-plasticity' dilemma [16].

This is only a sketch of an argument that a *memory-centric*, SDR-based, hierarchical system can retain the ability to learn new information, both episodic and semantic, throughout essentially unbounded lifetimes, without suffering CF, and while retaining fixed response time for learning and best-match retrieval. Quantitaively analyzing this overall argument is a current major research goal.

References

[1] Aljundi, Rahaf, Babiloni, Francesca, Elhoseiny, Mohamed, Rohrbach, Marcus, & Tuytelaars, Tinne. 2017. Memory Aware Synapses: Learning what (not) to forget. *CoRR*, **abs/1711.09601**.

[2] Barkat, T.R., Polley, D.B., & Hensch, T.K. 2011. A critical period for auditory thalamocortical activity". Nature Neuroscience. *Nature Neurosci.*, **14**(9), 1189–1196.

[3] Bengio, Emmanuel, Bacon, Pierre-Luc, Pineau, Joelle, & Precup, Doina. 2015. Conditional Computation in Neural Networks for faster models. *ArXiv e-prints*, Nov., arXiv:1511.06297.

[4] Blakemore, Colin, & Van Sluyters, Richard C. 1974. Reversal of the physiological effects of monocular deprivation in kittens: further evidence for a sensitive period. *The Journal of Physiology*, **237**(1), 195–216.

[5] Cheetham, Claire E., & Belluscio, Leonardo. 2014. An Olfactory Critical Period. *Science*, **344**(6180), 157–158.

[6] Daw, N. W., & Wyatt, H. J. 1976. Kittens reared in a unidirectional environment: evidence for a critical period. *The Journal of Physiology*, **257**(1), 155–170.

[7] Erzurumlu, Reha S., & Gaspar, Patricia. 2012. Development and Critical Period Plasticity of the Barrel Cortex. *The European Journal of Neuroscience*, **35**(10), 1540–1553.

[8] Foster, David J. 2017. Replay Comes of Age. *Annual Review of Neuroscience*, **40**(1), 581–602.

[9] French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, **3**(4), 128–135.

[10] French, Robert. 1994. Dynamically Constraining Connectionist Networks to Produce Distributed, Orthogonal Representations to Reduce Catastrophic Interference. *Pages 335–340 of: In Proceedings of the 16th Annual Cognitive Science Society Conference*. Erlbaum.

[11] Friedmann, Naama, & Rusou, Dana. 2015. Critical period for first language: the crucial role of language input during the first year of life. *Current Opinion in Neurobiology*, **35**, 27–34.

[12] Fusi, Stefano, Drew, Patrick J., & Abbott, L. F. 2005. Cascade Models of Synaptically Stored Memories. *Neuron*, **45**(4), 599–611.

[13] Fusi, Stefano, Miller, Earl K., & Rigotti, Mattia. 2016. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, **37**, 66–74.

[14] Graves, Alex, Wayne, Greg, & Danihelka, Ivo. 2014. Neural Turing Machines. *ArXiv e-prints*, Oct., arXiv:1410.5401.

[15] Graves, Alex, Wayne, Greg, Reynolds, Malcolm, Harley, Tim, Danihelka, Ivo, Grabska-Barwińska, Agnieszka, Colmenarejo, Sergio Gómez, Grefenstette, Edward, Ramalho, Tiago, Agapiou, John, Badia, Adrià Puigdomènech, Hermann, Karl Moritz, Zwols, Yori, Ostrovski, Georg, Cain, Adam, King, Helen, Summerfield, Christopher, Blunsom, Phil, Kavukcuoglu, Koray, & Hassabis, Demis. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, **538**, 471.

[16] Grossberg, S. 1980. How does a brain build a cognitive code? *Psychological Review*, **87**(1), 1–51.

[17] Hintzman, Douglas L. 1984. MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers*, **16**(2), 96–101.

[18] Hubel, D. H., & Wiesel, T. N. 1970. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of Physiology*, **206**(2), 419–436.

[19] Indyk, Piotr, & Motwani, Rajeev. 1998. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. *Pages 604–613 of: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. STOC '98. New York, NY, USA: ACM.

[20] Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil, Veness, Joel, Desjardins, Guillaume, Rusu, Andrei A., Milan, Kieran, Quan, John, Ramalho, Tiago, Grabska-Barwinska, Agnieszka, Hassabis, Demis, Clopath, Claudia, Kumaran, Dharshan, & Hadsell, Raia. 2017. Overcoming catastrophic forgetting in neural networks. *PNAS*, **114**(13), 3521–3526.

[21] Kraska, Tim, Beutel, Alex, Chi, Ed H., Dean, Jeffrey, & Polyzotis, Neoklis. 2017. The Case for Learned Index Structures. *ArXiv e-prints*, Dec., arXiv:1712.01208.

[22] Kumaran, Dharshan, Hassabis, Demis, & McClelland, James L. 2016. What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, **20**(7), 512–534.

[23] 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. *Psychol. Rev.*, **102**, 419–457.

[24] McCloskey, M., & Cohen, N. J. 1989. *Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem*. Vol. 24. Academic Press. Pages 109–165.

[25] Muir, Darwin W., & Mitchell, Donald E. 1975. Behavioral deficits in cats following early selected visual exposure to contours of a single orientation. *Brain Research*, **85**(3), 459–477.

[26] Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

[27] Pettigrew, J. D., & Freeman, R. D. 1973. Visual Experience without Lines: Effect on Developing Cortical Neurons. *Science*, **182**(4112), 599–601.

[28] Pickett, Marc, Al-Rfou, Rami, Shao, Louis, & Tar, Chris. 2016. A Growing Long-term Episodic & Semantic Memory. *ArXiv e-prints*, Oct., arXiv:1610.06402.

[29] Poggio, T., Mutch, J., Leibo, J. Z., Rosasco, L., & Taccheti, A. 2012. *The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work)*. Report TR-2012-035. MIT CSAIL.

[30] Poo, Cindy, & Isaacson, Jeffry S. 2007. An Early Critical Period for Long-Term Plasticity and Structural Modification of Sensory Synapses in Olfactory Cortex. *J. Neurosci.*, **27**(28), 7553–7558.

[31] Pritzel, Alexander, Uria, Benigno, Srinivasan, Sriram, Badia, Adrià Puigdomènech, Vinyals, Oriol, Hassabis, Demis, Wierstra, Daan, & Blundell, Charles. 2017. Neural Episodic Control. *Pages 2827–2836 of:* Precup, Doina, & Teh, Yee Whye (eds), *Proceedings of the 34th International Conference on Machine Learning*. Proceedings of Machine Learning Research, vol. 70. International Convention Centre, Sydney, Australia: PMLR.

[32] Rinkus, Gerard. 1996. *A Combinatorial Neural Network Exhibiting Episodic and Semantic Memory Properties for Spatio-Temporal Patterns*. Thesis.

[33] Rinkus, Gerard. 2010. A cortical sparse distributed coding model linking mini- and macrocolumn-scale functionality. *Frontiers in Neuroanatomy*, **4**.

[34] Rinkus, Gerard. 2017. A Radically New Theory of how the Brain Represents and Computes with Probabilities. *arXiv preprint arXiv:1701.07879*.

[35] Rinkus, Gerard J. 2014. Sparsey^{TM}: event recognition via deep hierarchical sparse distributed codes. *Frontiers in Computational Neuroscience*, **8**(160).

[36] Rust, Nicole C., & DiCarlo, James J. 2010. Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *The Journal of Neuroscience*, **30**(39), 12978–12995.

[37] Sabour, Sara, Frosst, Nicholas, & E Hinton, Geoffrey. 2017. Dynamic Routing Between Capsules. *ArXiv e-prints*, Oct., arXiv:1710.09829.

[38] Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, & Salakhutdinov, Ruslan. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**(1), 1929–1958.

[39] Wang, J., Liu, W., Kumar, S., & Chang, S. F. 2016. Learning to Hash for Indexing Big Data - A Survey. *Proceedings of the IEEE*, **104**(1), 34–57.

[40] Weston, Jason, Chopra, Sumit, & Bordes, Antoine. 2014. Memory Networks. *ArXiv e-prints*, Oct., arXiv:1410.3916.

[41] Zoccolan, Davide, Kouh, Minjoon, Poggio, Tomaso, & DiCarlo, James J. 2007. Trade-Off between Object Selectivity and Tolerance in Monkey Inferotemporal Cortex. *J. Neurosci.*, **27**(45), 12292–12307.

4 Supplementary Material

		Equation	Short Description
	1	$u(i) = \sum_{j \in \mathrm{RF_U}} x(j) w(j,i)$ $h(i) = \sum_{j \in \mathrm{RF_H}} x(j, t-1) w(j,i)$	Compute raw U (and H, if applicable) input sums.
	2	$U(i) = \; u(i) / \pi_U \, w_{\max}$ $H(i) = h(i) / \pi_H \, w_{\max}$	Compute normalized U (and H, if applicable) input sums. In this paper's simulations, $\pi_U = 12$ and $\pi_H = Q - 1$.
	3	$V(i) = \begin{cases} H(i)^{\lambda_H} U(i)^{\lambda_U} & t \geq 1 \\ U(i)^{\lambda_U} & t = 0 \end{cases}$	Compute local evidential support for each cell. In this paper, $\lambda_H = \lambda_U = 1$, unless otherwise stated.
Hard Max →	4	$\hat{V}_j = \max_{i \in C_j} \{V(i)\}$	Find the max V, $\hat{V}_j$, in each CM, C_j
Modulate Transfer Fn.	5	$G = \sum_{q=1}^{Q} \hat{V}_q / Q$	Compute G as average $\hat{V}$ value over the Q CMs
	6	$\eta = 1 + \left(\left[\frac{G - G^-}{1 - G^-} \right]^+ \right)^{\gamma} \times \chi \times K$	Determine expansivity (η) of V-to-η sigmoid function. In this paper, γ=2, χ=100, $G^- = 0.1$
	7	$\sigma_1 = \frac{\left((\eta - 1)/0.001\right)^{1/\sigma_4} - 1}{e^{\sigma_2 \sigma_3}}$	Sets σ_1 so that the overall sigmoid shape is preserved over full η range. $\sigma_2 = 7.0, \sigma_3 = 0.4, \sigma_4 = 9.5$
	8	$\mu(i) = \frac{(\eta - 1)}{\left(1 + \sigma_1 e^{-\sigma_2 (V(i) - \sigma_3)}\right)^{\sigma_4}} + 1$	To each cell, apply sigmoid function, which collapses to constant fn, $\mu(i) = 1$, when $G \leq G^-$)
Soft Max	9	$\rho(i) = \frac{\mu(i)}{\sum_{k \in CM} \mu(k)}$	In each CM, normalize relative (μ) to final (ρ) probabilities of winning
	10	Select a final winner in each CM according to the ρ distribution in that CM, i.e., soft max.	

Figure 2: A simple non-hierarchical, variant of Sparsey's Code Selection Algorithm (CSA), involving only the combination of bottom-up (U) and horizontal (H) signals, see [33] for details. The H matrix carries signals (recurrently) from the previously active (at T-1) SDR code(s). More general, hierarchical variants are given in [34, 35]. One can readily see by inspection, that the the algorithm does not iterate over stored items. Its dominant step is Step 2, requiring a single iteration over afferent weights, for all units, i.e., input summations. There is one feedforward pass through the steps. In the case of a hierarchical model with many macs on many levels, on each time step (e.g., frame of an input sequence), there is a single upward processing pass through all the levels, in which, generally, a subset of the macs meet activation criteria and execute the CSA, thus assigning/storing codes (or in the retrieval case, activating the best-matching stored codes). The details of how the CSA preserves similarity are given in [32–35]. A Java app (http://www.sparsey.com/CSA_explainer_app_page.html) is available of a slightly simplified version of the CSA, allowing experimentation with parameters affecting the learned simlarity-preserving mapping.

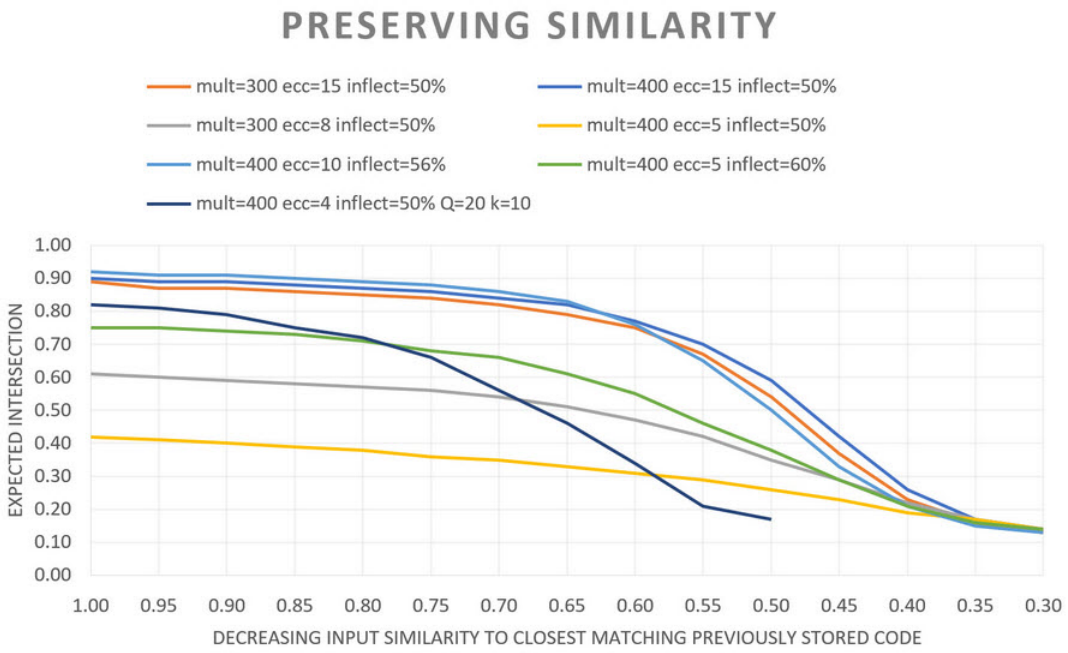

Figure 3: Simulation results of a simplified CSA and a model with only bottom-up (U) inputs from an input level to a single SDR coding field, consisting of Q=10 WTA CMs (in one case, Q=20), each with K=10 binary units, showing that spatial input similarlity (pixel-wise overlap, x-axis decreases towards right) is statistically preserved into size of intersection of SDR codes (y-axis). These results are produced from Java app mentioned in previous caption (http://www.sparsey.com/CSA_explainer_app_page.html). The various curves correspond to different settings of some of the CSA parameters controlling the shape and size of the sigmoid transfer function. "mult" corresponds to parameter χ in Step 6 of Fig. 2. "ecc" and "inflect" have their appropriate meaning viz. a sigmoid function, but don't correspond precisely to the σ parameters in Fig. 2. The main point of this figure is simply to show that the fixed-time CSA statistically preserves similarity from pixel overlap to SDR intersection, and does so in a smoothly graded way.

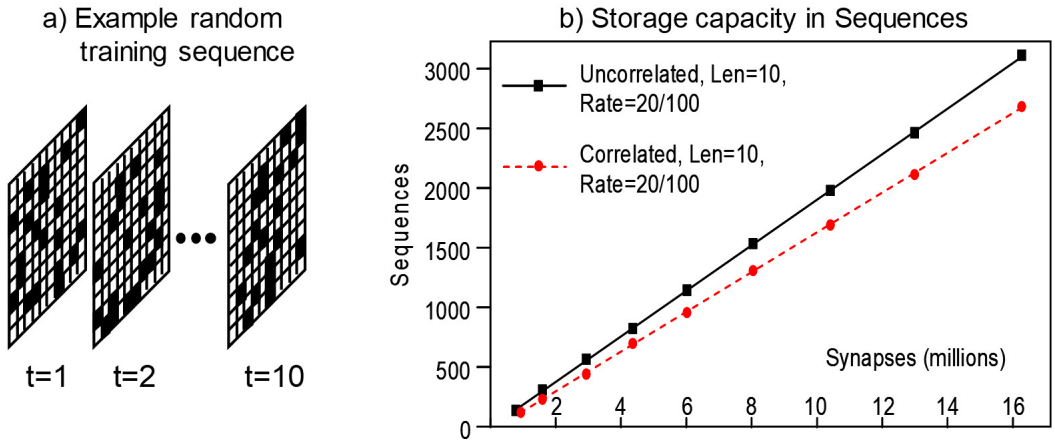

Figure 4: a) Example synthetic data sequence (10 frames, 20 out of 100 randomly chosen features per frame) for testing model capacity. b) Capacity is linear in the weights. "Uncorrelated" denotes randomly created frames (20 out of 100 features, chosen independently on each frame). "Correlated" denotes the complex sequence case: an *alphabet* of a 100 frames is pre-created. Sequences are then created by choosing 10 frames from lexicon randomly with replacement). The model had Q=100 CMs, each with K=40 units, and approximately 16M weights. As the chart shows, over 3000 such uncorrelated sequences were learned/stored with one trial each, while permitting an average retrieval accuracy of average 97%. See [32] and http://www.sparsey.com/Sparsey_Storage_Capacity.html for more details. Also see [32] for other experiments demonstrating the ability to learn extremely long time dependences.

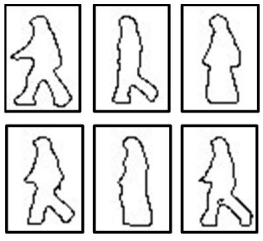

Figure 5: Example showing the pre-processing applied to spatial inputs, e.g., MNIST images, or frames of video. We edge filter, binarize, and skeletonize the inputs, all extremely cheap, local operations. These are a few frames from a Weizmann video.

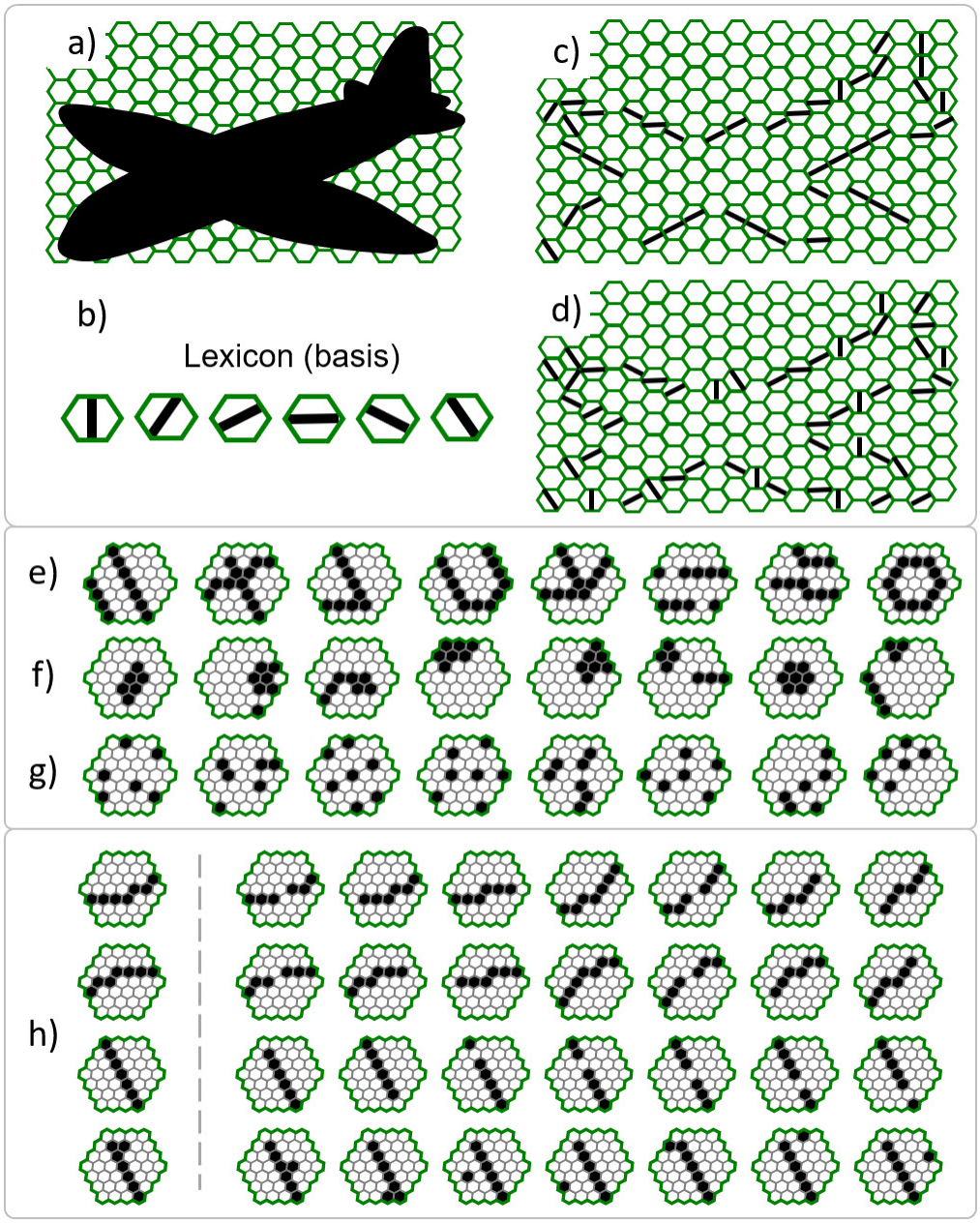

Figure 6: Explanation of why the pre-processing (e.g., Fig. 5) and a further constraint on the number of active features in a mac's bottom-up RF, in combination with the recursive, compositional (part-whole) structure of natural objects, greatly constrain the size of the basis needed for a mac to adequately represent its input space, i.e., represent all future inputs to its RF with sufficient fidelity to support expected future tasks (both probabilistic inference tasks and classifications), in particular, tasks defined at higher spatial/temporal scales, which will depend, in a complex way, on the fidelities of all macs at all levels of the hierarchy. For a-d, see text. e) Examples of inputs that are statistically plausible, but have too many active features, and so do not activate the mac and and are not stored. f) Examples of inputs that are statistically plausible, but will not survive the preprocessing, and so will not be stored. g) Examples of inputs that will survive the preprocessing but are statistically very unlikely (i.e., unlikely to be due to structural regularities), and so will not be stored. h) At left of each row is a statistically plausible input with an acceptable number of active features, which will thus be stored (assigned an SDR) if presented. Examples to right of dashed line have varying degrees of pixel overlap with the leftmost pattern, but would presumably be well-represented by the leftmost pattern, and thus not need to be explicitly stored, supporting the adequacy of a smaller basis.