A stable memory scaffold with heteroassociative learning produces a content-addressable memory continuum

Summary: Long-term memory is content addressable, in that partial cues are sufficient to drive recognition or recall of complete objects and events. Several content addressable memory (CAM) architectures have been proposed to model long-term memory, including the Hopfield network [1], several variants of the Hopfield network [2, 3, 4] and overparameterized autoencoders [5]. However, all of these architectures exhibit a memory cliff, beyond which adding a single pattern leads to catastrophic loss of all patterns. Here we propose a novel and biologically motivated memory architecture, Memory through Scaffolded Heteroassociation (MESH), that generates a CAM continuum: storage of information-dense patterns up to a critical capacity results in complete recovery of all patterns and storage of a larger number of patterns results in partial reconstruction of the stored patterns. This partial reconstruction continues up to an exponentially large number of patterns resulting in correct recognition of each of the stored patterns. Inspired by the entorhinal-hippocampal circuit, MESH contains a bipartite attractor network that stores a large dictionary of well-separated fixed points that serve as a pre-defined "memory scaffold". Arbitrary dense patterns are then stored by associating them to the pre-defined scaffold states. This novel combination of predefined attractor states along with heteroassociative learning that hooks patterns on to scaffolding states results in a biologically plausible CAM continuum, that approaches the theoretical upper-bound on information storage in neural networks [6, 7]. We believe that this is the first model of a content-addressable memory that automatically trades off pattern number and pattern richness; it makes the testable prediction that biological memory systems may exploit pre-existing scaffolds to acquire new memories, potentially consistent with the preplay of hippocampal sequences before they are used for representing new environments [8].

Further Details: Theoretical results suggest that the amount of information that can be stored and recovered in a CAM is bounded by the total number of synapses in the network [6, 7]. This defines a memory continuum, wherein the storage capacity of a CAM is inversely proportional to the information per pattern that can be recovered by the network, assuming that the number of synapses are held constant. Existing CAM models lie only as discrete points on this continuum (Fig. 1a), wherein they successfully store a total amount of information that is of the order of the number of synapses — however, storage of any additional information in these networks results in a catastrophic breakdown of the memory, with complete loss of all stored information (Fig. 1b). This is evidenced by the rapid drop off beyond a critical capacity in the mutual information between the recovered patterns and the stored patterns in each of the memory architectures considered in Fig. 1b — the Hopfield network, a classical CAM model in neuroscience, has a critical capacity of $0.14N$ beyond which all stored information is lost [9]; the pseudoinverse-learning variant of the Hopfield network [2] has a critical capacity of $N/2$; sparse Hopfield networks [3] store patterns with lesser information ($\sim Np\log p$ bits of information per pattern) and have larger critical capacities of $\mathcal{O}(N/(p\log p))$; and, overparametrized autoencoders that implement associative memory [5] present a critical capacity dependent on the number of hidden units. In each of these networks storage of marginally greater amounts of information results in loss of all stored information.

Overcoming this major drawback, we aim to construct a memory architecture that implements the entire extent of the CAM continuum (Fig. 1c), such that *without any change of architecture*, one can store increasingly larger magnitudes of information while only affecting the quality of information recovered in a continuously varying fashion. We present a novel architecture, MESH (Fig. 1d), that contains N_l-dimensional random binary k-sparse labels projected randomly to a layer of dense N_h-dimensional binary hidden states. Back projections from the hidden layer to the label layer are learnt through pairwise hebbian learning. We model the label layer to approximate attractor dynamics by enforcing a global k-hot code. Arbitrary N_f-dimensional dense patterns in the feature layer are associated with the hidden states through online pseudo-inverse learning [10]. Given noisy feature layer states, the network can perfectly reconstruct N_h number of patterns; and partially reconstruct more than N_h patterns spanning a memory continuum (Fig. 1e). The mutual information curve is asymptotically proportional to the theoretical upper bound of total information bounded by the number of synapses (shown as the dashed line). Correspondingly, Fig. 1f demonstrates that the information rate (information per synapse) in MESH asymptotically approaches a constant. This is in sharp contrast to existing memory models that rapidly drop to zero information rates beyond their critical capacities.

The interaction between the label and hidden layers implements attractor dynamics with an exponential capacity (Fig. 1g) that stores a set of predefined k-hot labels and corresponding hidden states. When presented with highly corrupted versions of hidden states, this two-layer attractor is able to recall all the stored labels and hidden states with high accuracy (provided that the number of hidden neurons is larger than a critical value N_h^{crit}), as can be seen in Fig. 1h. This allows a large number of feature states to then be "tagged" to each of the stored hidden states. Given corrupted versions of these stored features, the network is able to recover all the corresponding hidden states and label states perfectly (recognition). Further, it can perfectly reconstruct the feature states up to N_h patterns (Fig. 1i), proportional to the theoretical upper bound for perfect reconstruction. Storage of additional patterns up to an exponentially large number of patterns results in successful recovery of partial pattern information (Fig. 1e). In all cases, the recovered features remain within the nearest-neighbor basin of the originally stored pattern (black curve in Fig. 1i).

We thus provide a plausible model for biological memory, with continuous flexible control on the number of patterns stored and information recovered per pattern. Our results suggest that future studies may reveal the existence of a predefined memory scaffold in neural circuitry associated with long-term memory, whose signatures may also be seen in behavioral studies quantifying the inverse relationship between rate of decay of information as a function of the number of stored memories.

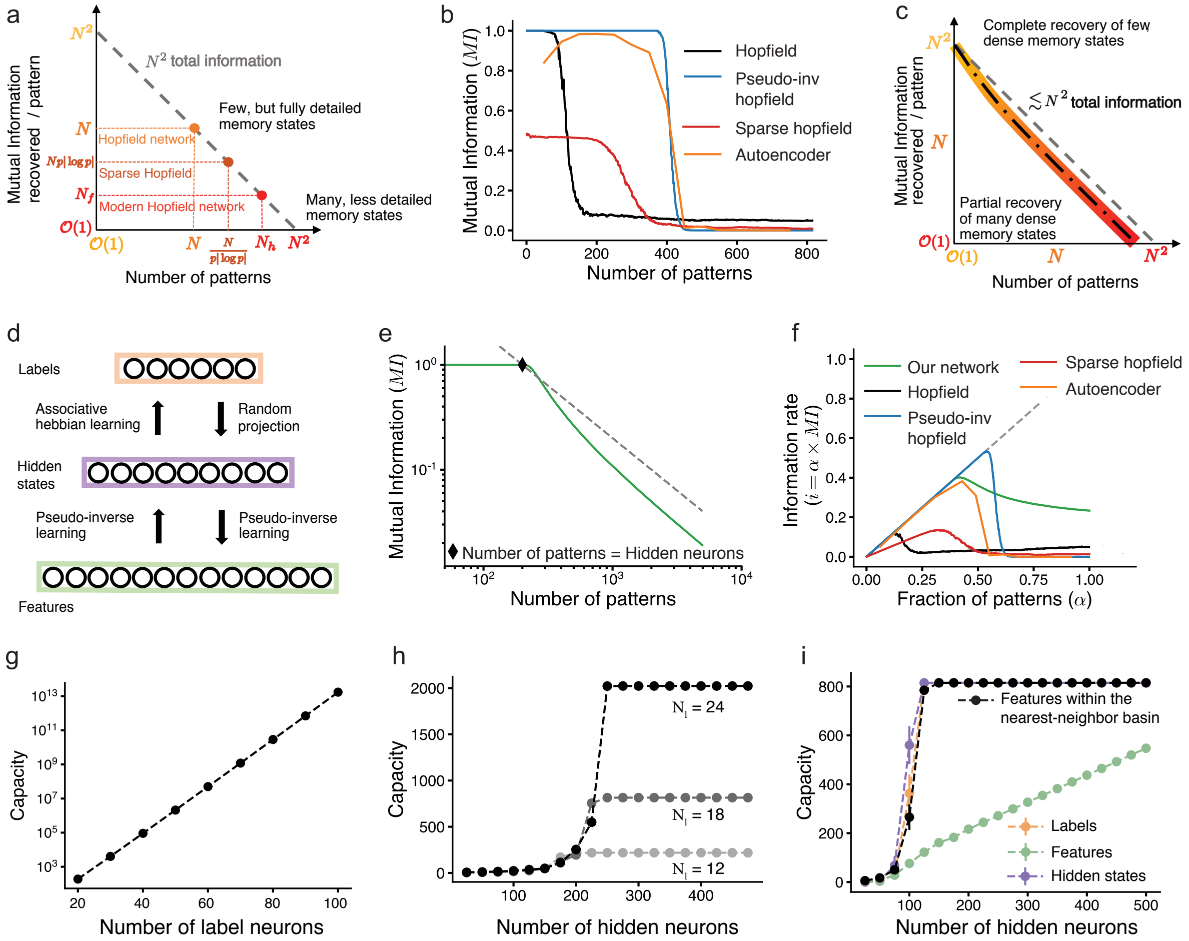

Figure 1: (a) Theoretical upper-bound on information storage for networks with N^2 synapses, and where existing networks lie relative to it. (b) Mutual information (per bit per pattern) between the stored and the recovered pattern as a function of number of patterns stored in existing networks: standard Hopfield network, Hopfield network trained with pseudo-inverse learning rule, sparse Hopfield network, and overparameterized autoencoder. (All networks chosen to have $\approx 5 \times 10^5$ synapses) (c) Desired CAM continuum. (d) Our proposed model, MESH. (e) Mutual information (per bit per pattern) in our network as a function of number of patterns stored in the network. (f) Comparison of information rate across different networks relative to our model. α is defined as the fraction of patterns relative to the number of feature neurons in each network. (g) Exponential capacity of the label-hidden layer attractor network as a function of the number of label neurons assuming a constant sparsity (90%) of stored labels. (h) Capacity of label-hidden layer attractor network as a function of the number of hidden neurons (computed with 20% input noise injected in the hidden layer, and allowing up to 3% recovery error). Different curves show the capacity corresponding to different sizes of the label layer for labels with a constant number of active bits ($k = 3$). (i) Capacity of the our network for storing arbitrary patterns (5% input noise perturbation in feature layer, capacity presented for perfect pattern recovery). Here $N_l = 18$, $k = 3$ and $N_f = 816$.

[1] Hopfield. (1982). [2] Personnaz et al. (1985). [3] Tsodyks & Feigel'man. (1988). [4] Krotov & Hopfield. (2021). [5] Radhakrishnan et al. (2020). [6] Abu-Mostafa. (1989). [7] Gardner. (1988). [8] Dragoi & Tonegawa (2011). [9] Amit et al. (1987). [10] Tapson & van Schaik. (2013).