Recurrent Neural Network for "Aha! That's it!"

Hiro-Fumi Yanai* and Daisuke Senga
Ibaraki University, Department of Media and Telecommunications
Hitachi, Ibaraki, Japan.
*hfy@ieee.org

In retrieving memories, how we humans know it is right or not? For years since neural network models for memory storage and retrieval (i.e., associative memory) are vastly and deeply studied in 1980's, researchers sought the ways for increasing the memory capacity and enlarging radii of attraction of memories. In those studies, mechanisms that determine the correctness of retrieved memories are rarely discussed. An example of studies that explicitly deals with this cognitive property of humans ("Aha! That's it!") in the context of neural networks is the one proposed in Ref. [1]. The model exhibits implicit check information whether the retrieved state is correct or not through elaborate and careful choices of network designs. We propose a fairly simple model that is a natural extension of standard associative memories, and that can check by itself whether the retrieved state is right or not. Our model is natural because it utilizes the correlation matrix (kind of Hebbian synapses), and extension of the model is done just by implementing feedback loops computable all the way from the values of the correlation matrix.

To be specific, the state of the network x_t (a vector with components ± 1) evolves as

$$x_{t+1} = \mathrm{sgn}\left(W x_t - s x_t\right).$$

The matrix W is defined by the correlation matrix W_c of embedded memories as $W = W_c \sum_{k=0}^{\tau} (I - \alpha W_c)^k$, with parameters α and τ, where I is the unit matrix. The term with the scalar s represents a negative feedback to each neuron itself, which plays important role in our model. It is easy to rewrite this seemingly nonlocal retrieval processes by combinations of local rules [2]. Also, the matrix W bridge the gap between the correlation matrix ($\tau = 0$) and the orthogonal projection matrix ($\tau \to \infty$) [2].

Conceptual picture of non-Aha and "Aha!" is shown in Fig. 1. The initial states which are incorrect but share good proportion of features with the embedded memory converge to it when the radius of attraction, ρ ($0 \leq \rho < 0.5$), is large enough (non-Aha), and of course the embedded memory remains unchanged. If ρ is small enough, incorrect states are rejected, so that, virtually, only the embedded memory itself remains unchanged—We call this feature the "Aha!". We can easily control ρ with the magnitude of self-connection s. For example, if $r = 0.3$ (r is the ratio of the number of embedded memories to the number of neurons), $\rho \simeq 0.2$ for $s \simeq 0.4$, $\rho \simeq 0.1$ for $s \simeq 0.5$, and $\rho \simeq 0.0$ for $s \simeq 0.7$ (Fig. 2).

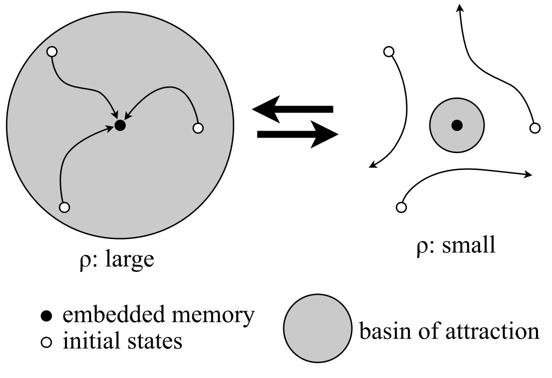

Figure 1: Conceptual picture of non-Aha (left) and "Aha!" (right).

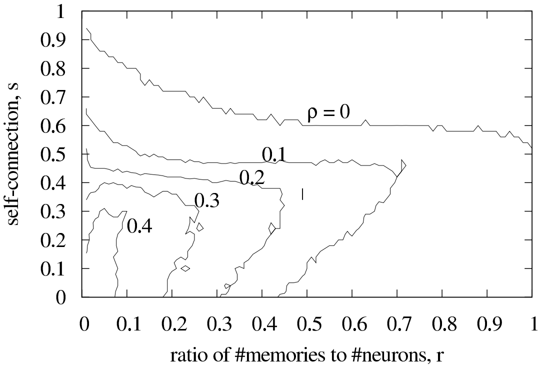

Figure 2: Contour plot of radius ρ of attraction of randomly assigned embedded memories, where $\#\text{neurons} = 500, \alpha = 0.5$ and $\tau = 5$.

References

[1] Hopfield, J.J., Searching for memories, Sudoku, implicit check bits, and the iterative use of not-always-correct rapid neural computation, Neural Computation, vol.20, pp.1119–1164 (2008)
[2] Yanai, H.-F., Equivalence and differences in recall and storage dynamics of associative memory, Proceedings of the International Conference on Neural Information Processing (ICONIP) 96, pp.581–586 (1996)