Large Associative Memory Problem in Neurobiology and Machine Learning

Dmitry Krotov
MIT-IBM Watson AI Lab
IBM Research

John Hopfield
Princeton Neuroscience Institute
Princeton University

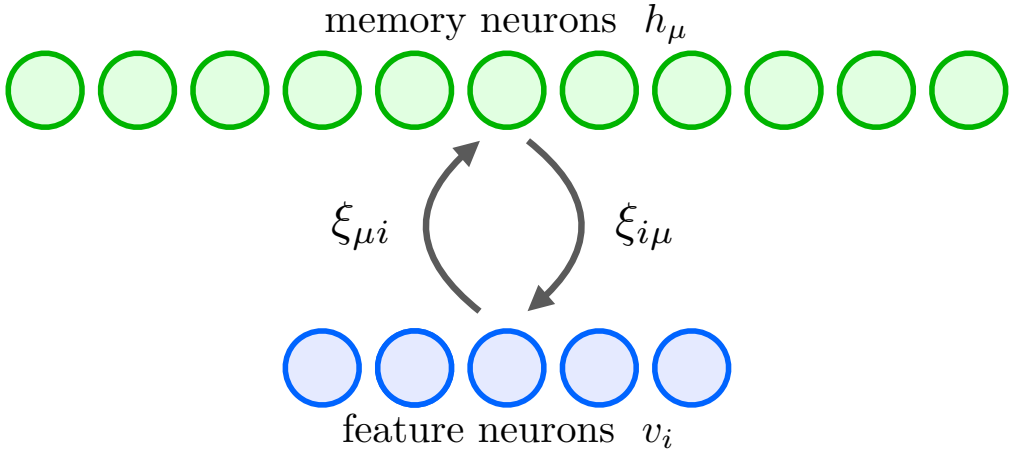
MIT-IBM Watson AI Lab

ICLR 2021

PRINCETON Neuroscience INSTITUTE

Microscopic theory of modern Hopfield networks or Dense Associative Memories

$$\begin{cases} \tau_f \frac{dv_i}{dt} = \sum_{\mu=1}^{N_h} \xi_{i\mu} f_\mu \quad v_i + I_i \\ \tau_h \frac{dh}{dt} = \sum_{i=1}^{N_f} \xi_{\mu i} g_i \quad h_\mu \end{cases}$$



$$E(t) = \Big[\sum_{i=1}^{N_f} (v_i \quad I_i) g_i \quad L_v \Big] + \Big[\sum_{\mu=1}^{N_h} h_\mu f_\mu \quad L_h \Big] \quad \sum_{\mu,i} f_\mu \xi_{\mu i} g_i$$

Modern Hopfield networks have a very large memory storage capacity

Standard Associative Memory classical Hopfield network

$$E = -\sum_{i,j=1}^{N_f} \sigma_i T_{ij} \sigma_j \qquad T_{ij} = \sum_{\mu=1}^{N_h} \xi_{\mu i} \xi_{\mu j}$$

σ_i - dynamical variables

$\xi_{\mu i}$ - memorized patterns

N_f - number of feature neurons

N_{mem} - number of memories

$$E = -\sum_{\mu=1}^{N_h} \Big(\sum_{i=1}^{N_f} \xi_{\mu i} \sigma_i \Big)^2$$

$$N_{\text{mem}} \approx 0.14 N_f$$

Dense Associative Memory modern Hopfield network

$$E = -\sum_{\mu=1}^{N_h} F\Big(\sum_i \xi_{\mu i} \sigma_i \Big)$$

$F(x) = x^n, \quad \text{with} \quad n \geq 2$

$F(x) = \exp(x)$

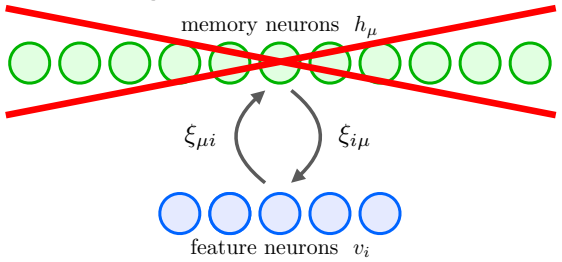
$$N_{\text{mem}} \approx \min(N_f^{n-1}, N_h)$$

$$N_{\text{mem}} \approx \min\left[\exp(\alpha N_f), N_h \right]$$

$$\alpha < \ln(2)/2$$

- Krotov & Hopfield "Dense Associative Memory for pattern recognition" NeurIPS 2016
- Demircigil et al. "On a model of associative memory with huge storage capacity" Journal of Statistical Physics 2017
- Ramsauer et al. "Hopfield networks is all you need" ICLR 2021

Effective theory for feature neurons

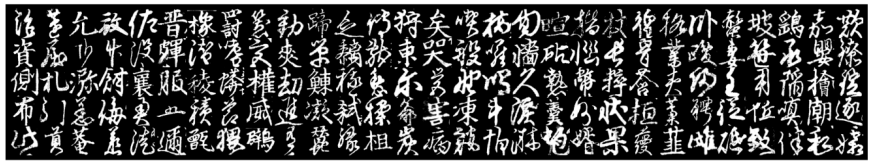


	Model A	Model B	Model C
Lagrangian functions	$L_h = \sum_\mu F(h_\mu)$ $L_v = \sum_i \lvert v_i \rvert$	$L_h = \log\left(\sum_\mu e^{h_\mu}\right)$ $L_v = \frac{1}{2}\sum_i v_i^2$	$L_h = \sum_\mu F(h_\mu)$ $L_v = \sqrt{\sum_i v_i^2}$
energy	$E = -\sum_{\mu=1}^{N_h} F\left(\sum_i \xi_{\mu i}\sigma_i\right)$	$E = \frac{1}{2}\sum_{i=1}^{N_f} v_i^2 - \log\left(\sum_\mu \exp(\sum_i \xi_{\mu i} v_i)\right)$	$E = -\sum_\mu F\left(\sum_i \xi_{\mu i}\frac{v_i}{\sqrt{\sum_j v_j^2}}\right)$
effective update rule	$\tau_f \frac{dv_i}{dt} = \sum_{\mu=1}^{N_h} \xi_{i\mu} f\left(\sum_{j=1}^{N_f} \xi_{\mu j}\sigma_j\right) - v_i$	$\tau_f \frac{dv_i}{dt} = \sum_{\mu=1}^{N_h} \xi_{i\mu}\mathrm{softmax}\left(\sum_{j=1}^{N_f} \xi_{\mu j} v_j\right) - v_i$	$\tau_f \frac{dv_i}{dt} = \sum_\mu \xi_{i\mu} f\left[\sum_j \xi_{\mu j}\frac{v_j}{\sqrt{\sum_k v_k^2}}\right] - v_i$

- Dense Associative Memory for pattern recognition
- On a model of associative memory with huge storage capacity
- Hopfield networks is all you need

Some examples of problems in AI and biology

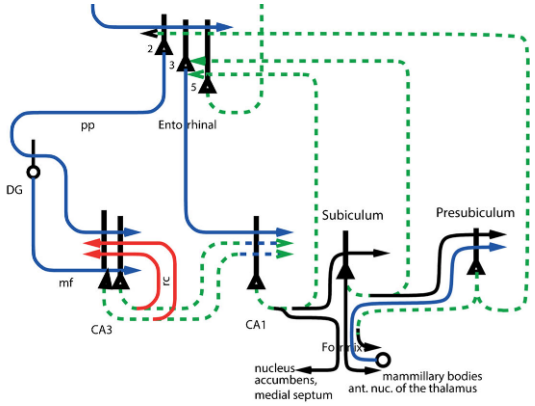
Pattern memorization. Consider a small gray scale image 64x64 pixels. If one treats the intensity of each pixel as an input to a feature neuron the standard Hopfield network would be able to only memorize approximately 573 distinct patters (0.14 times 4096). Yet, the number of all possible patterns of this size that one can imagine is far bigger. For instance, Kuzushiji-Kanji dataset includes over 140,000 characters representing 3832 classes with most of the characters recognizable by humans. A well educated Japanese person can recognize about 3000-5000 character classes, which means that those classes are represented in his/her memory. In addition, for many characters a person would be able to complete it if only a portion of that character is shown.



Clanuwat, et al., “Deep learning for classical Japanese literature”

Another area of the hippocampus potentially related to the mathematical model described in this paper is the area CA1, which, in addition to receiving inputs from CA3, also receives inputs directly from the entorhinal cortex, and projects back to it. In this interpretation pyramidal cells of the CA1 would be interpreted as the memory neurons in our mathematical model, while the cells in the layer III of the entorhinal cortex would be the feature neurons. The feedback projections from CA1 go primarily to layer V of the entorhinal cortex, but there are also projections to layers II and III. While it is possible to connect the proposed mathematical model of Dense Associative Memory with existing networks in the hippocampus, it is important to emphasize that the hippocampus is involved in many tasks, for example imagining the future, and not only in retrieving the memories about the past. For this reason it is difficult at present to separate the network motifs responsible for memory retrievals from the circuitry required for other functions.

Cortical-hippocampal system. The hippocampus has long been hypothesised to be responsible for formation and retrieval of associative memories. One candidate for associative memory network in the hippocampus is the CA3 area, which consists of a large population of pyramidal neurons in conjunction with an inhibitory network that keeps the firing rates under control. There is a substantial recurrent connectivity among the pyramidal neurons, which is necessary for an associative memory network. There are also several classes of responses of those neurons in behaving animals, one class being place cells. In addition to place cells it contains other neurons that do not respond in experiments designed to drive place cells, but presumably are useful for other tasks. One possible way of connecting the mathematical model proposed in this paper with the existing anatomical network in the brain is to assume that some of the pyramidal cells in CA3 correspond to the feature neurons in our model, while the remaining pyramidal cells are the memory neurons. For example, place cells are believed to emerge as a result of aggregating inputs from the grid cells and environmental features, e.g. landmark objects, environment boundaries, visual and olfactory cues, etc. Thus, it is tempting to think about them as memory neurons (which aggregate information from feature neurons to form a stable memory) in the proposed model.



Rolls, E.T. The storage and recall of memories in the hippocampo-cortical system. Cell Tissue Res 373, 577–604 (2018)

Conclusions

- We have proposed a general microscopic theory of Dense Associative Memories or modern Hopfield networks that on one hand has a large memory storage capacity (much greater than the number of input neurons), and, at the same time, is manifestly describable in terms of only two-body synapses.
- This theory is described by an energy function, which decreases on the dynamical trajectories, and a system of non-linear differential equations with fixed point attractors.
- Excluding degrees of freedom associated with hidden neurons from this theory results in many effective models previously discussed in the literature, such as conventional Hopfield networks with binary and continuous states, Dense Associative Memories with binary states, modern Hopfield networks for the feature neurons, attention mechanism, as well as some new models.
- We discuss potential applications in machine learning and neuroscience in which it might be useful to have a recurrent neural network capable of pattern completion with guaranteed convergence properties to the fixed points and a large number of those fixed points.