Hopfield Networks

An overview of the paper and ensuing online discussion

DL Theory Reading Group

January 4, 2023

Summary: A system of simple components interacting can display emergent behavior. For instance, out of the firing of a group of neurons emerges: perception, cognition, memory etc. The Hopfield network was a proposed recurrent neural network model with emergent associative memory. The network can recall memorised states given a noisy or incomplete starting point.

1 Introduction

Does computation emerge as a consequence of the behavior of large groups of neurons? This paper proposes a recurrent neural network model that aims to display emergent phenomena:

- Time stability of memory
- Generalisation
- Time-sequential Memory

An aim of the modelling here was for the details of the model to be largely irrelevant, in the same way that for sound wave generation collisions of particles are necessary, but any (sensible) inter-atomic force gives appropriate collisions for sound waves to be generated.

There are physical systems whose behavior can be used as content addressable memory. [1] Such a physical system would be described by general coordinates, where a particular coordinate describes an instantaneous condition of the system.

There are many such systems, but to be useful as memory it must flow towards locally stable points, like a body of water with multiple whirl pools. Anything close to a particular whirlpool gets pulled in.

The system has coordinates:

$$X = (x_1, x_2, ..., x_N) \tag{1}$$

and locally stable points:

$$X_A, X_B, ... \tag{2}$$

where:

[1] Recall of memory initiated by comparing input pattern to memorised pattern.

$$X = X_A + \mathit{\Delta} \implies X \to X_A \tag{3}$$

In particular we want such a system [2] where any particular state can be made into a locally stable state.

1.1 Hopfield Network

The network consists of a set of neurons $\{V_i\}_i^N$ where $V_i = 1$ means neuron i is on and $V_i = 0$ means i is off. The strength of the connection between neuron i and j is T_{ij}, with $T_{ij} = 0$ denoting the absence of a connection between i and j. The system is describe by a vector of N bits.

States change according to these rules:

$$\begin{matrix} V_i \to 1 \\ V_i \to 0 \end{matrix} \text{ if } \sum_{j \neq i} T_{ij} V_j \begin{matrix} > U_i \\ < U_i \end{matrix} \tag{4}$$

Where U_i is the threshold for neuron i, set to 0 unless otherwise stated.

These updates occur randomly in time and asynchronously across neurons.

1.1.1 Difference between this and perceptron

This network differs from a perceptron in three ways:

- Perceptron only has forward connections, the emergent behavior comes from having backwards and forwards connections
- "Perceptron studies usually made a random net of neurons deal directly with a real physical world and did not ask the questions essential to finding the more abstract emergent computational properties"
- Perceptrons have synced neurons, all the updates happen at the same time for all neurons.

1.2 Memory Storage

We want to store a series of states $\{V^s\}_1^n$. We set:

$$T_{ij} = \sum_s (2V_i^s - 1)(2V_j^s - 1) \tag{5}$$

With $T_{ii} = 0$, from this we get:

$$\sum_j T_{ij} V_j^{s'} = \sum_s (2V_i^s - 1) \left[\sum_j (2V_j^s - 1) \right] \equiv H_j^{s'} \tag{6}$$

The mean value of the bracketed term is 0 unless $s = s'$, for which the mean is $N/2$. This psuedoorthogonality gives:

$$\sum_j T_{ij} V_j^{s'} = \langle H_j^{s'} \rangle \approx (2V_i^{s'} - 1)\frac{N}{2} \tag{7}$$

Which is positive if $V_i^{s'} = 1$ and negative if $V_i^{s'} = 0$. Aside from noise from terms with $s \neq s'$ the stored state should be a stable point of the system.

They claim:

[2] Such a system could be continuous, or discrete like the original Hopfield Network

> Such matrices T_{ij} have been used in theories of linear associative nets to produce an output pattern from a paired input stimulus, $S1 \to O1$. A second association $S2 \to O2$ can be simultaneously stored in the same network. But the confusing simulus $0.6S1 + 0.4S2$ will produce a generally meaningless mixed output $0.6O1 + 0.4O2$ Our model, in contrast, will use its strong nonlinearity to make choices, produce categories, and regenerate information and, with high probability, will generate the output $O1$ from such a confusing mixed stimulus.

The authors note that Hebbian learning is capable of producing T_{ij} as above.

1.3 Collective behavior of the model

To see there are stable limit points we show that there is an energy function associated with this model that decreases as the model updates.

First assume $T_{ij} = T_{ji}$[3] and define the following energy:

$$E = -\frac{1}{2} \sum_{i \neq j} \sum T_{ij} V_i V_j \tag{8}$$

Then the change in energy due to a change in V_i is:

$$\Delta E = -\frac{1}{2} \Delta V_i \sum_{i \neq j} T_{ij} V_j \tag{9}$$

Which is always negative[4] so E is monotonically decreasing, which indicates stable limit points.

This is isomorphic to an Ising spin glass model with symmetric couplings, for which it is known there many locally stable states[5].

For non-symmetric models showing that locally stable limit points exist is non-trivial [6], but some justification is given:

> Why should stable limit points or regions persist when $T_{ij} \neq T_{ji}$? If the algorithm at some time changes V_i from 0 to 1 or vice versa, the change of the energy defined in 8 can be split into two terms, one of which is always negative. The second is identical if T_{ij} is symmetric and is "stochastic" with mean 0 if T_{ij} and T_{ji} are randomly chosen. The algorithm for $T_{ij} \neq T_{ji}$ therefore changes E in a fashion similar to the way E would change in time for a symmetric T_{ij} but with an algorithm corresponding to a finite temperature.

1.3.1 Simulation test of stability

To investigate the models behaviour they run Monte Carlo simulation of the asymmetric model with number of neurons $N = 30$ [7]

They found:

- The system would not ergodically[8] wander through the state space
- The system would settle into limiting behaviors, the most common being a stable state
- 50 trials with a particular random T would result in 2-3 end states:
  - The system would settle into stable states, a few end states collected all flow from the initial state space

[3] Without this assumption the change in energy can be decomposed into two terms, one which is always negative and another which is the same if $T_{ij} = T_{ji}$ and stochastic with mean 0 if not

[4] If the change in V_i is positive then $V_i = 1$ so by 7 $\sum_{i \neq j} T_{ij} V_j$ is negative and thus their product is negative. The reverse holds also.

[5] Scott Kirkpatrick and David Sherrington. Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384, 1978

[6] Tianping Chen and Shun Ichi Amari. Stability of asymmetric hopfield networks. *IEEE Transactions on Neural Networks*, 12 (1):159–163, 2001

[7] They also run experiments with 100 neurons, but were stymied by computational issues.

[8] Visiting all states

  - Or a simple cycle might occur $A \to B \to A \to ...$
  - Chaotic wandering in a small region of the state space, but with the wandering happening within a small Hamming distance of a particular state.

The upshot is: even if $T_{ij} \neq T_{ji}$ this could act as a physical addressable memory.

1.3.2 Simulation test of memory

The system is started at each nominally "remembered state" and the system is allowed to run forward. This is what they found:

- About $0.15N$ states can be "remembered"
- For a fixed number of neurons, as the amount of memories required to be remembered increased, the number of states that evolved to stable states decreased
- The amount of reliably recoverable memories increases with the number of neurons
- Given arbitrary starting points 85% end in assigned memories, 10% end in stable states with no obvious meaning, 5% end in stable states near assigned memories Corrupting known memories: if the Hamming distance between a corrupted state and its assigned memory is ≤ 5 then the state evolved to its nearest memory 90% of the time, reducing to 20% of the time if the distance is ≤ 12

The authors describe it as:

> The phase space flow is apparently dominated by attractors which are the nominally assigned memories, each of which dominates a substantial region around it. The flow is not entirely deterministic, and the system responds to an ambiguous starting state by a statistical choice between the memory states it most resembles.

- New memories can be added to T_{ij} but adding more than capacity causes all memory states to be irretrievable unless there is a way of forgetting.
- With $N = 100$ a hamming distance of about 50 ± 5 is required for memories to be treated as distinct with certainty.

References

Tianping Chen and Shun Ichi Amari. Stability of asymmetric hopfield networks. *IEEE Transactions on Neural Networks*, 12(1):159–163, 2001.

Scott Kirkpatrick and David Sherrington. Infinite-ranged models of spin-glasses. *Physical Review B*, 17(11):4384, 1978.