# Hopfiled model to recall MNIST digits

Tizayi Zirereza[1] and Panagiotis Georgiou[1]

[1] *School of Physics and Astronomy, University of Nottingham, Nottingham, NG7 2RD, UK*

The MNIST dataset was reduced to only include pictures with labels "3", "6", "1", "0" and "4" and a Hopfield model was used to store 5 patterns from pictures in the dataset. The model was then tested upon retrieving the patterns from random pictures in the dataset. The model achieved an overall recollection accuracy of 80 ± 10%. The accuracy for retrieving the digits "3","6","1","0" and "4" were 81 ± 2%, 76 ± 1%, 98 ± 1%, 76 ± 2% and 69 ± 1% respectively.

## I. INTRODUCTION

An associative memory model has the ability to recall a stored variable when presented with a similar input variable or an input with partial information. Therefore, such models provide a way to efficiently deal with problems such as recovering damaged images. This work implements the Hopfield model which is a simple RNN inspired by the ferromagnetic behavior of spin variables in a lattice in an attempt to recognise hand written digits from the MNIST data set. The idea of using the physics of the Ising spin model to create a computational addressable memory was introduced in 1982 by J. J. Hopfield [2] following the work of W. A. Little who used the Ising model ideas to theorize how the brain produces memories [3].

## II. THEORY

The model is comprised of a lattice of N spin variables each of which can take a binary value of either {-1,1}. The set of the N spins variables is defined by the vector $\vec{S} = \{S_1, S_2, ...S_N\}$. The Hopfield model stores information in the form of patterns which are specific configurations of the spin variables.

The overlap is a measure of how close a given configuration $\vec{S}$ is to a given pattern $\vec{\xi^{\mu}}$:

$$m_\mu(\vec{S}) = \frac{1}{N}\vec{S} \cdot \vec{\xi^{\mu}}. \tag{1}$$

### A. Pseudo inverse prescription

The model's dynamics are determined by the Hamiltonian of the spin system. The standard Hopfield Hamiltonian, suits problems in which the patterns are almost orthogonal. However, in cases that the overlap of stored patterns is significant, the model often retrieves states that look nothing like the stored patterns. These states are called spurious and they correspond to global minima of the free energy function. So, in order to define a model that is more robust with non-orthogonal stored patterns, the pseudo-inverse Hamiltonian was used. [1].

The pseudo inverse Hamiltonian is obtained by defining a $(p \times p)$ matrix Q

$$Q_{\mu,\nu} = \frac{1}{N}\vec{\xi_\mu} \cdot \vec{\xi_\nu} \tag{2}$$

where $\mu$ and $\nu$ are indices which run over all possible combinations of the patterns encoded in the model. This matrix can then be used to calculate the weights $W_{ij}$ of the model:

$$W_{ij} = \frac{1}{N}\sum_{\mu\nu} \xi_i^\mu (Q^{-1})_{\mu\nu}\xi_j^\nu \tag{3}$$

thus the Hamiltonian function is

$$H = -\frac{1}{2}\sum_{i \neq j} W_{ij}S_iS_j. \tag{4}$$

This method increases the storage capacity of the model and thus guarantees the stability of a larger set of stored patterns [1].

### B. Monte Carlo (Metropolis) algorithm

For the retrieval of a pattern, the system's Hamiltonian is minimised using the Monte Carlo metropolis algorithm. Spins chosen randomly from the lattice are inverted if by doing so, it causes the system's energy to decrease. Otherwise, the inversions are accepted with a probability $p = e^{-\frac{\Delta E}{k_B T}}$ where $\Delta E$ is the change in energy, $T$ is the temperature of the system, and $k_B$ is the Boltzmann constant. The Temperature T is responsible for the influence of randomness in the Monte Carlo dynamics of the energy minimization. Above a critical temperature the model is governed by random fluctuations and is not able to reach a global minimum. Given enough random iterations below the critical temperature, the model will reach an equilibrium at a minimum of the free energy function which should correspond to a stored pattern's spin configuration.

**Algorithm 1** Monte Carlo metropolis algorithm

```
for Number of iterations do
    choose random spin to try and flip
    compute change in energy $\Delta E$ of the proposed flip
    if $\Delta E \leq 0$ then
        Accept flip proposal
    else
        Generate random value $x$ between 0 and 1
        if $x \leq e^{-\frac{\Delta E}{k_B T}}$ then
            Accept flip proposal
        end if
    end if
end for
```

## III PROCEDURE

The MNIST digits are grey-scale images of size (28 × 28). The pixels with values 0 were converted to -1 whereas the rest were scaled down to +1 in order to match the binary spin nature of the Hopfield model. The selection of patterns was performed via the method described in III.A . Due to the inevitable non-orthogonality of the patterns the pseudo-inverse Hamiltonian was chosen to describe the model. To deduce the critical temperature, the model was run using 2 stored patterns of digits "0" and "4". The starting configuration chosen was random noise. Then, the temperature was varied between 0 and 2 and the overlap with the pattern "4" was calculated. This procedure was repeated 100 times for each temperature and the overlap was averaged. As T had to be smaller than the critical temperature of 0.5, as deduced from figure 1, so its value was arbitrarily chosen to be 0.2.

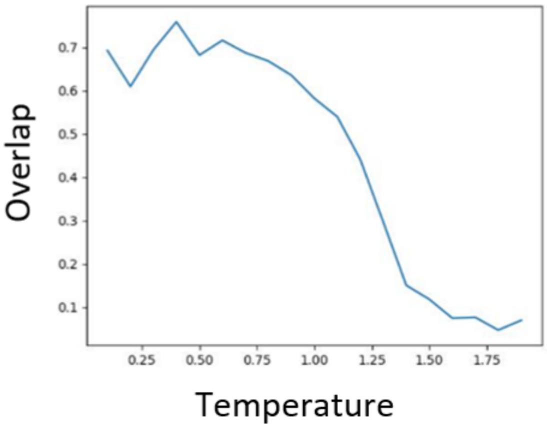

FIG. 1. A graph of the overlap of a randomly created input with the pattern "4" against Temperature is displayed. The patterns used to run the model were "4" and "0".

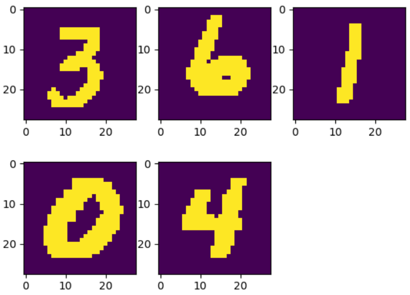

FIG. 2. The 5 patterns stored in the model

### A. Pattern selection

The patterns to be stored within the model were selected from the MNIST dataset such that the overlap between the the stored digit and other digits of the same label is maximised while the overlap between the stored digit and digits of a different label is minimised. To achieve this, 20 random examples of each digit were sampled from MNIST a loss was calculated for each digit

$$L(P) = \sum_{j}^{9M} (P \cdot D_j)^2 - M \sum_{i}^{M-1} (P \cdot U_i)^2 \qquad (5)$$

where P is the pattern in question, D is the set of patterns with a different label to P, U is the set of patterns with the same label as P, and M = 20 is the number of random samples for each digit. An example of each digit was obtained by finding the sample digit which had the lowest loss defined in equation 5.

### B. Testing performance

The patterns for the numbers "3","6","1","0" and "4" were selected from this optimised sample to test the models performance. These patterns are shown in figure 2. To test the performance of our model to successfully recall patterns, the model was run using 20 sweeps of N/2 iterations of algorithm 1 and was tested on 100 new examples of each digit randomly selected from the dataset. It was decided that an example would be described as successfully recalled if its overlap with the stored pattern of the same label is larger than 0.99. The test was repeated 10 times for each digit.

## IV RESULTS

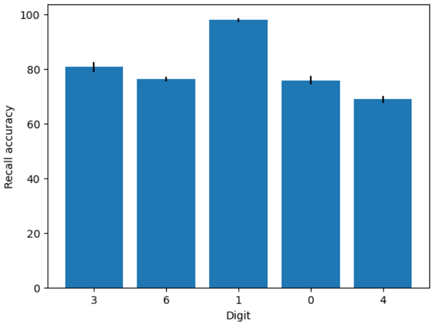

FIG. 3. The average accuracy achieved for each digit with the black bars being the standard deviation of the accuracy values

The model was able to recall on average $80 \pm 10\%$ of digits correctly. The model performance for each digit is shown in figure 3. The model showed the best performance on recalling the digit ”1” as the model was able to recall $98 \pm 1\%$ of the test digits with label ”1”. The worst performance was seen on the digit ”4” for which the model was only able to achieve a recollection accuracy of $69 \pm 1\%$ on test digits with label ”4”. We hypothesise that the reason for such poor performance on the digit ”4” can be attributed to the larger variety of shapes the digit ”4” took in the MNIST data set, while generally the digit ”1” is very consistent in its shape.

## V CONCLUSION

The ability of a Hopfield model, defined by a pseudo-inverse Hamiltonian, to store and retrieve handwritten digits of “0”, “1”, “3”, “4” and “6” from the MNIST dataset was investigated. The pictures from the dataset selected to become the patterns of those numbers were chosen according an optimization function that ensured maximum overlap with pictures of the same number and minimum overlap of pictures of different number. The accuracy of the model retrieving its patterns from other pictures of the dataset classified with the same digit were: $81 \pm 2\%$, $76 \pm 1\%$, $98 \pm 1\%$ and $76 \pm 2\%$ and $69 \pm 1\%$ for the digits “3”, “6”,”1”,”0” and ”4” respectively, while its overall accuracy was calculated to be $80 \pm 10\%$. The pattern of “1” was retrieved with the highest accuracy while the pattern of ”4” was retrieved with the lowest accuracy. These results can be explained by the assumption that ”1” was the pattern with the most consistent shape while ”4” was the one with the least. To test our hypothesis, future work could look into using the loss function described in equation 5 to check the consistency of shapes in the MNIST dataset.

## REFERENCES

[1] D. J. Amit. *Modeling brain function: the world of attractor neural networks / Daniel J. Amit.* eng. Cambridge University Press, 1989, pp. 172–174. ISBN: 978-0-521-42124-9.

[2] J J Hopfield. “Neural networks and physical systems with emergent collective computational abilities.” In: *Proceedings of the National Academy of Sciences of the United States of America* 79.8 (Apr. 1982), pp. 2554–2558. ISSN: 0027-8424.

[3] W. A. Little. “The existence of persistent states in the brain”. en. In: *Mathematical Biosciences* 19.1 (Feb. 1974), pp. 101–120. ISSN: 0025-5564. DOI: `10.1016/0025-5564(74)90031-5`.