

Associative Memories with "Killed" Neurons: the Methods of Recovery^{*}

A.M. Reznik, A.S. Sitchov, O.K. Dekhtyarenko, and D.W. Nowicki

The Institute of the Mathematical Machines and Systems, Ukrainian National Academy of Science

03187, 42 Glushkov Str, Kiev, Ukraine

neuro@immsp.kiev.ua

Abstract—We consider re-learning ability of a Hopfield-type network after killing some neurons. Neurons were "killed" by means of nullification of corresponding rows and columns of the synaptic matrix. We show that one can restore recognition ability of this network using re-training with the vectors, which were memorized before. The number of vectors needed is equal to the number of deleted neurons. It does not depend on network's size and on volume of stored data.

INTRODUCTION

It is known that a classical Hopfield network has decreasing convergence ability with respect to number of memorized vectors. Such a network cannot store more vectors than 14% of neurons' number [1]. Pseudoinverse learning rule enables to increase this ratio up to 25% [2]. In this case we must store exact weight values in the synaptic matrix (at least 7 bits per weight, [3]). But sometimes disturbance of accurate weight values does not decrease convergence ability of the Hopfield-type network. On the contrary, some distortions may make it work better as an associative memory. Let us note some examples of such "useful distortions": methods of desaturation [4], which allows increasing of the memorize ability about 5 times, pseudoinverse adaptive filter [5], method of weight selection (it enables to reduce number of weights to 30% of original quantity not worsening associative-memory capabilities, [6]). These examples illustrate the effect of information redundancy inherent to Hopfield-type associative memories

Therefore the following question seems to be interesting: Does the redundancy effect work if some neurons of the Hopfield-type network are completely destroyed? Could one recover the associative memory in this case? To answer these questions we consider a pseudo-inverse network with some neurons "killed" by means of nullification of all their synaptic weights (both for the inputs and the outputs). Once exposed to such a distortion, the network loses its ability to converge, i.e. the destruction of associative memory takes place and all its content becomes inaccessible. We show that it is possible to recover the network completely via retraining it with some of the previously stored vectors. The number of vectors needed is equal to the number of deleted neurons. It does not depend on network's size and on volume of stored data. This phenomenon looks like recovery of amnesia patients after reminding them significant events of their past.

THE MODEL OF NEURAL NETWORK

According to the J. Hopfield's scheme each neuron is connected to each other and itself; forward and backward connections have the same weight. Weighted sum of output signals forms a postsynaptic potential (PSP). Depending on a sign of this sum the neuron's output possess the values $+1$ or -1 . The outputs form a N -dimensional vector of current network's state, and weights form a synaptic matrix C . Under certain conditions for this matrix the network has stable states called attractors. If a network state isn't an attractor the process of convergence will take place. Convergence ends in a nearest attractor. The convergence process looks like an associative recall, that's why networks of this type are known as associative memories. The algorithm for calculation of synaptic matrix from given set of attractors [2] is based on solution of the stable state equation:

$$CU = U. \quad (1)$$

Here U is a matrix $M \times N$. The vectors of desired attractor states are its columns. The solution of this equation has a form:

$$C = UU^+, \quad (2)$$

Besides the main attractors defined by equation (1) the network has spurious equilibria defined as solutions of non-linear stability equation.

If there are many spurious attractors the network may stop at them before converging to main attractors. Spurious attractors exert influence only if $M/N > 1/10$. For slightly saturated networks they could be neglected. The pseudoinverse algorithm is commonly used to compute value of the matrix C . It enables to successively update the matrix in memorizing of each vector from U [4].

$$C^m = C^{m-1} + \Delta^m; \\ \Delta^m = (I - C^{m-1})UU^T(I - C^{m-1}) / U^T(I - C^{m-1})U, \quad (3)$$

where C^m is a value of synaptic matrix after memorizing M vectors.

An expression for Δ^m could be rewritten in the form:

$$\Delta_{ij}^m = (u_i - s_i)(u_j - s_j)/q, \quad (4)$$

^{*}This research was supported by INTAS-01-0257

where $s_i = \sum_{j=1}^N C_{ij} u_j$ is a postsynaptic potential of j -th neuron; $q = \sum_{j=1}^N (u_j - s_j) u_j$

A normed value of q is called coefficient of distinction. It describes a component of vector U orthogonal to m -1 previously stored vectors.

$$k = q / N. \quad (5)$$

The value of k decreases as U approaches to linear hull of U_i and reaches zero when U belongs to this linear hull. Such a behavior resembles the resonance effect as it emphasizes the difference between similar stored vectors.

Quality of associative memories could be also described in terms of attraction radius. Direct attraction radius is the largest value of Hamming distance between initial point and the attractor for that the examination procedure starting at this point will still converge to the corresponding attractor during one iteration.

The value of direct attraction radius depends on data nature and could be estimated by the following expression [4]:

$$H < 0.5(N-1)^{0.5} [1 - (1 + \alpha)M/N] [M/N - (M/N)^2]^{-0.5} + 1 \quad (6)$$

A positive value $\alpha < 1$ is called desaturation coefficient. Diagonal terms of the matrix should be multiplied by this value to fasten convergence and to resist saturation effects.

During experimentation we use the concept of the full attraction radius, which is the maximum Hamming distance covered by the network as it reaches the state of attraction (regardless of the number of iterations). The full attraction radius normally is 3-5 times greater than attraction radius given by (6) and is more important for the practical applications.

CHANGE OF ATTRACTION PROPERTIES AFTER REMOVING CONNECTIONS

Partial removal of connections causes change of the PSP. Small changes do not affect its sign; so they don't cause network's state changes. In this case the effect could be seen due to change of the distinction coefficient. This coefficient describes network's behavior in the neighborhood of the attractor.

The histograms of k are for the network of 256 neurons displayed in the fig. 1. In this network all connections of n (randomly selected) neurons were destroyed after memorizing M vectors. For these experiments the NeuroLand software package [6] was used. Memorized and test vectors were generated randomly. We can see that the distribution of k looks like normal. The removal of just 10% of neurons leads to the significant non-zero k values that signifies the loss of the synaptic matrix projective properties and the deterioration of the network recall capacities.

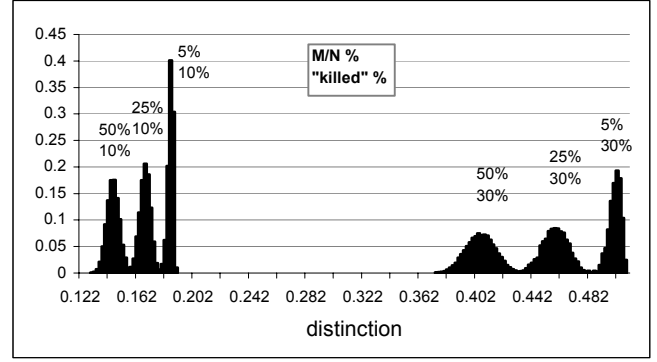


Fig. 1. The k value distinction for various memory saturations (M/N) and "killed" neuron portions.

TRAINING DISTORTED NEURAL NETWORK

In the fig. 1 we can see resonance properties of the NN changed. These changes increase a probability of stop in the spurious attractors. So the volume of data recalled is decreased. Moreover, loss of projectivity of the synaptic matrix may damage memorizing recall abilities more seriously. If one uses formulae (3, 4) to correct perturbed matrix then distortion will be only cumulated. That does finally completely destroy the associative memory. To verify these assumptions we have made some experiments. Under certain conditions the results were completely opposite.

The results for memorizing 120 vectors and "killing" 40 neurons in the network of 256 neurons are presented in the figs. 2 and 3. For examination we used desaturation with $\alpha = 0.1$. According to formula (6) $H = 8.7$, and the complete attraction radius was about 35. To obtain the AR experimentally we use noisy values of stored vectors as initial conditions for the convergence process. Complete attraction radius was defined as a maximum value of H for that network was converged to the correct image for 95 of 100 noise instances.

Data for "killing" 40 neurons and re-memorizing 10, 30, and 40 vectors are displayed in the fig. 2, 3. Vectors for retrain were randomly selected from images initially stored. Fig 2 depicts the values of attraction radius for the remaining part of the network (216 neurons).

We can see that the destroyed network completely losses its convergence ability. Only for one of 120 initially stored images (#100) attraction radius was non-zero. During additional training attraction radius is recovered completely for retrain images and partially for the rest. As number of retrain vectors gets equal to number of "killed" neurons attraction properties of the network are completely restored. The distinction coefficient with respect to retrain dynamics is shown the figure 3. Unlike fig. 2 this graph takes into account all 256 neurons of the network. Dynamics of k with respect to retrain looks like attraction radius dependencies. After re-memorizing 40 vectors it turns to zero for all stored images.

The recovery effect for associative memories takes place if

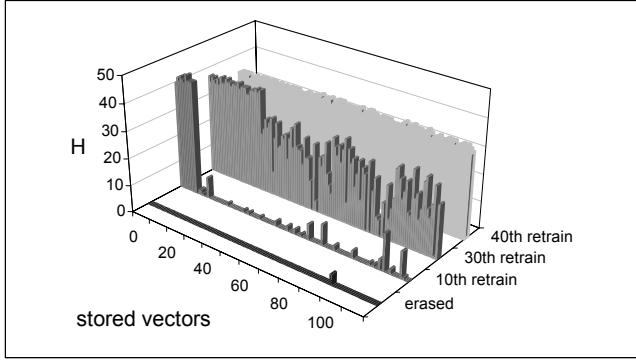


Fig. 2. The changes of attraction radius with net retrain.

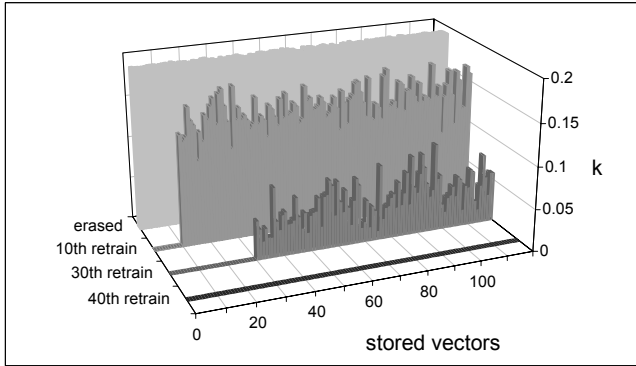


Fig. 3. The changes of distinction coefficient with net retrain

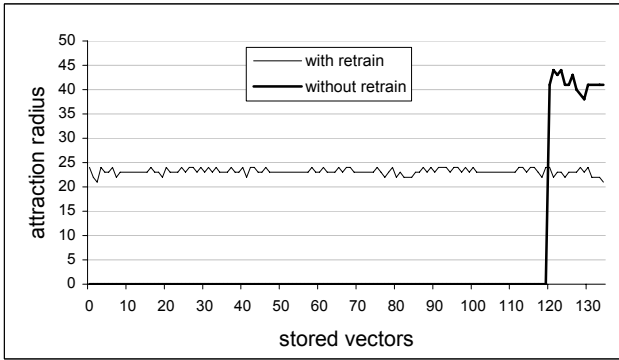


Fig. 4. The change of attraction radius with net extra train

and only if retrain data belongs to a set of initial images. If we take different vectors attraction abilities of the network could not be restored. That follows from fig. 4: attraction radius after storing 15 new vectors in the network. These are vectors #121 and more. The dependencies for new training that is done immediately after killing 40 neurons (bold line) or following retraining (thin line) are displayed in this graph. We can see that without retraining attraction radius is zero for all previously stored vectors. For the new stored vectors attraction radius (it was more than 40) increases in comparison with the value before damaging network. It could be explained by the fact that the trace of the synaptic matrix was decreased by the neurons removal, thus the sum of

eigenvalues defining the associative memory saturation became smaller.

SPECTRA OF DISTORTED SYNAPTIC MATRICES

The recovery process of associative memory could be observed by examining spectra of synaptic matrices. Data for the matrix 256×256 with 120 stored vectors after nullification of 40 rows and columns are displayed in fig 5. There are graphs of sorted spectra before retraining and after 10, 30, and 40 retrain vectors. We can see that "killing" neurons preserves rank of the matrix, although values of 40 eigenvalues of 120 strongly decreased. After storing each following vector rank of the matrix increases by one, an eigenvector with value of 1 is added, 40 weakened eigenvalues are changing. After retrain they form a tail of the spectrum; these eigenvalues are less than 0.47. There is an orthogonal component with respect to original memorized vectors in eigenvectors of the corrupted matrix.

Fig. 5 displays data for dense filling of the memory ($M/N \geq 0.47$) after removal of 15% of neurons. Fig. 6 displays data for slight filling of the same network ($M/N \geq 0.15$) after removal of 10 neurons only. There are spectra before retraining, after 5 and 10 retrain vectors and after additional training with 20 new vectors.

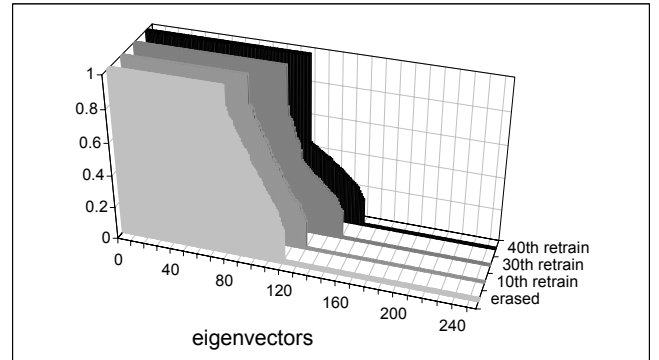


Fig. 5. Spectra evolution during the net retrain ($M = 120$)

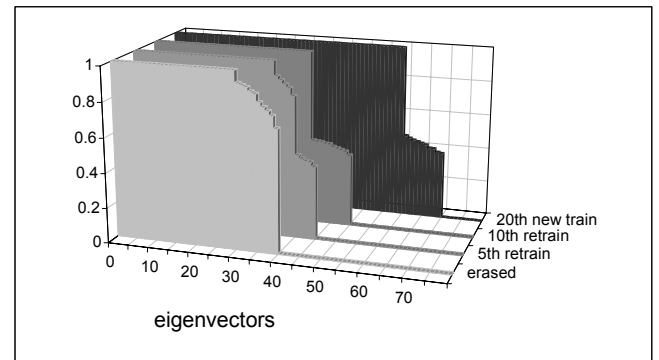


Fig. 6. Spectra evolution during net retrain and extra train with new data ($M = 40+20$)

Note that 20 new eigenvectors with $\lambda = 1$ almost do not influence on 10 small eigenvalues (appeared after corruption).

For slightly saturated memory non-projectivity of the matrix does not practically influence on the network's behavior. However for strongly filled memory there are many spurious attractors; so, available associative memory is reduced. We can depress spurious attractors using desaturation. To reduce non-projectivity more drastically one could raise the synaptic matrix to high power; so only eigenvalues close to 1 would be preserved.

CONCLUSION

Considering the results of our experiments it is worth to emphasize that the discussed phenomenon covers not only a simple substitution of "killed" neurons and their connections, but also the complete recovery of the autoassociative memory functions for the remaining part of the net (its functionality was lost due to the loss of the network ability to converge towards attractors). Therefore it is possible to say about the recovery both of stored data and the autoassociative access mechanism.

We have done experiments on network recovery for Hopfield-type networks with 64 to 2048 neurons for different levels of saturation and corruption. In all the case the network could be completely recovered using retraining with original data.

The number of retrained images needed for the complete recovery was always equal to the number of deleted neurons. We have not revealed any differences caused by the various possible selections of retrained images from the already stored ones in the network. Such independence on the retrain set selection could be used for designing of ultra stable systems that are capable to preserve their features in spite of constant resource degradation. In order to achieve it, these systems must undergo recovery via retraining of a control dataset more intensively than the resource degradation occurs.

This phenomenon looks like recovery of amnesia patients after reminding them significant events of their past. So, we can make the conjecture that in the biological neural systems there are mechanisms working like pseudo-inverse associative memories. We can cite some facts of biological neuroscience to support this conjecture. Not that most nerve fibers are less than 2 mm long; size of the dendrite branching area (0.1-0.4 mm) has the same order. In 1 mm² there are thousands strongly connected neurons. So, they form local structures of 100-1000 neurons looking like Hopfield-type networks. We can suppose that these structures perform information memorizing and retrieval. In scope of our model and results long-term memory acts by systematic reactivation of local structures; they are re-training with previously memorized patterns. Memory corruption is prevented due to continuous conscious and subconscious work of the brain; also due to the dream activity. The further research in cooperation with specialists in biological neuroscience and psychologists is needed to verify these conjectures.

REFERENCES

- [1] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Nat. Acad. Sci.*, vol. 79, pp. 2554-2558, Apr. 1982.
- [2] L. Personnaz, I. Guyon, G. Dreyfus, "Collective computational properties of neural networks: New learning mechanisms," *Phys. Rev. A*, vol. 34, no. 5, pp. 4217-4228, 1986.
- [3] M. Weinfeld, "A fully digital integrated CMOS Hopfield network including learning algorithm," in *Proc. Int. Workshop WLSI Art. Intell.*, Univ. of Oxford, E1-E10, 1988.
- [4] A.M. Reznik, D.O. Gorodnichy, A.S. Sitchov, "Regulating feedback bond in neural networks with learning projectional algorithm," *Cybernetics and system analysis*, vol. 32, no. 6, pp. 868-875, 1996.
- [5] A.M. Reznik, "Non-iterative learning for neural networks," in *Proc. Int. Joint Conf. Neural Networks*, Washington, no. 548, July, 1999.
- [6] A.S. Sitchov, "Weight selection in neural networks with pseudoinverse learning rule," (in Russian) *Mathematical Machines and Systems*, vol.2, pp. 25-30, 1998.
- [7] A. M. Reznik, E.A. Kalina, A.S. Sitchov, E.G. Sadovaya, O.K. Dekhtyarenko, A.A. Galinskaya, "Multifunctional neurocomputer NeuroLand," (in Russian) in *Proc. Int. Conf. Inductive Simulation*, Lviv, Ukraine, vol. 1(4), pp. 82-88, May 2002.