ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 387-392

Nonlinear factorization in sparsely encoded Hopfield-like neural networks

A.M. Sirota, A.A. Frolov *, D. Husek **

Moscow Institute of Physics and Technology
*Institute of Higher Nervous Activity and Neurophysiology of the RAS
**Institute of Computer Science AS of the CR

The problem of binary factorization of complex patterns in recurrent Hopfield-like neural network was studied both theoretically and by means of computer simulation. The number and sparseness of factors mixed in patterns crucially determines the ability of an autoassociator to perform a factorization. Basing on experimental data on memory and learning one may suggest, that there exists a neural system of intermediate storage of information, which fulfills the function of binary factorization of the incoming polysensory information for its further effective storage in the form of elementary associatively bound factors. We suppose that field CA3 of the hippocampus possessing all properties of the autoassociative memory performs such function. This functional idea could be fruitfully applied to various memory related tasks (e.g. spatial navigation) and lead to some critical experiments.

Introduction

We define factorization as decomposition of a complex vector signal into a set of simple factors based on correlation between components of the former. The linear factorization in vector space (for example by means of Principal Component Analysis) is the simplest form of such decomposition.

The nonlinear case of factorization is obviously more sophisticated and can not be solved analytically. One particular form of nonlinear factorization is a binary one, where a complex vector signal (pattern) has a form of a logical sum of weighted binary factors: $\mathbf{X}=\bigvee_{l=1}^{L} \alpha_l \mathbf{f}^l$. This case of factorization allows an interpretation in terms of attractor neural networks with binary activity. The central idea is that network can easily learn cross-correlations that underlie in incoming complex pattern using Hebbian learning rule. Hebbian rule forms connections matrix as a covariance matrix for the set of learned patterns. Neurons that tend to fire together (represent one common factor) will be more correlated and corresponding connection strengths will be larger in respect to those neurons that belong to different factors. Hence, each group of neurons that forms a factor might correspond to the attractor of the network dynamics. Thus, in order to perform factorization one has to train the network with a set of complex patterns using correlational Hebbian learning rule. The procedure of factors extraction could be based on the search of attractors of the network dynamics

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 387-392

that might correspond to the factors. This paper is devoted to investigation of conditions under which factors actually form attractors in sparsely encoded Hopfield-like neural network [9] by means of computer simulation.

Model description

Detailed theoretical and computational analysis of sparsely encoded Hopfield-like neural networks was given elsewhere [1,2,6,7]. In contrast to these works we trained the network by set of complex patterns (nonlinear combinations of factors). In this sense usual learning procedure corresponds to the case of purely presented factors.

Thus, on the learning stage, fully connected network of N binary neurons was trained by a set of M patterns of the form $\mathbf{X}^m = \bigvee_{l=1}^{L} \alpha_l^m \mathbf{f}^l$, where $\mathbf{f}^l \in \mathrm{B}_p^N$ (pN is maintained constant) are L factors and $\alpha^m \in B_{p_f}^L$ are factor scores [1]. Both factors and factor scores were chosen statistically independent. Here p and p_f are sparsenesses (ratio of number of active elements to the total number of elements) of factors with respect to neurons and patterns with respect to factors. In a limit $p_f \to 0$ patterns become pure factors that correspond to ordinary Hopfield case. Connections matrix J was formed using the correlational Hebbian rule:

$$\mathbf{J}_{ij} = \frac{1}{A}\sum_{m=1}^{M}(X_i^m - q\{X^m\})(X_j^m - q\{X^m\}) \;\; i \neq j,\; \mathbf{J}_{ii} = 0\,, \quad (1)$$

where $q\{X^m\} = \sum_{i=1}^{N} X_i^m / N$ is the total activity of the pattern. As N increases the value of q approaches the expectation $M\{X_i^m\} = 1 - (1 - pp_f)^L$. Theoretical analysis showed that bias taken equal to expectation of remembered pattern activity provides the best separation between “false” and “true” modes (see below) on the first step of the evolution. Such form of bias corresponds to the biologically plausible global inhibition being proportional to overall neuron activity.

On the recall stage, on presentation of an initial pattern, the network was let to evolve until it stabilized in some attractor. The evolution of the network’s state is determined by the synchronous dynamics equation for activity X in time:

$$X_i(t+1) = \Theta(h_i(t) - T(t))\,,\; X_i(0) = f_i^l\,, i = 1,\ldots,\mathrm{N}, \quad (2)$$

where $$h_i(t) = \sum_{j=1}^{N} \mathbf{J}_{ij} X_j(t) \quad (3)$$

is synaptic excitation, Θ-step function, and $T(t)$ – activation threshold. The

[1] $\mathrm{B}_p^N = \{\mathbf{X} \mid X_i \in \{0,1\}, P\{X_i = 1\} = p \;\; \forall i = 1,\ldots,N\}$

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 387-392

threshold *T(t)* is chosen at each time step in such a way that sparseness of the network activity is kept constant and equal to *p*. Thus, on each step *n=pN* „winners" (neurons with greatest synaptic excitation) are chosen and only they are active on the next step. This procedure ensures that attractors are only point and cyclic of length two. The stable pattern (point attractor) or first pattern of cyclic attractor was taken as a resulting pattern of the recall process. In order to check whether factors do form attractors we took pure factors as initial network states.

Main parameters

For the analysis of information properties of a network some integral parameters are introduced. The recall quality is measured by overlap between initial and resulting vectors $m^f = m(X^{in}, X^f) = \frac{1}{N}\sum_{i=1}^{N} X_i^{in} X_i^f$. As a measure of the relative informational loading we use $\alpha \approx Lh(p)/N$, where $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the Shannon function. The informational capacity of the network is α_{cr} , which is a maximum α , for which stable states in the vicinities of stored factors still exist. If we define two more parameters: $C = Lp_f$ – *complexity* of patterns (average number of active factors in pattern) and $S = Np_f$, we shall obtain five parameters of the network: *N, M, C, S, p*. Then
$\alpha = Ch(p)/S$ (4).

The results of the simulation

In order to analyze the dependency of α_{cr} on parameters of the network the model was simulated on the computer. Calculations were performed for *N* from 200 to 4000, for *p*= 0.5, 0.1, 0.02. The program generated random factors, mixed them randomly (regarding sparsenesses *p* and p_f) into the set of *M* patterns, trained the net with this set, and tested the net with factors. On the basis of gathered statistics the distribution of final overlaps m^f was plotted. This distribution has two distinct modes: for $m^f \approx 1$ ("true") and $m^f \approx 0$ ("false"), that correspond to stabilization of the network in true and spurious attractor respectively. The threshold value $m^f{}_{thr}$ used for separation was determined as a point of a minimum in case of balanced distribution between two modes. The probability of existence of stable attractors in the vicinities of factors was estimated by the probability that m^f belongs to the „true" mode: $P = P\{m^f > m^f_{thr}\}$. It turned out, that as *M* increases these two modes tend to separate better, basins of attraction enhance and *P* increases, saturating at some M_{sat} ($\approx 10^3 / p_f$). The obtained values of *P* corresponding to different α were

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 387-392

approximated by the following function $P(\alpha) = 1/(1+\exp(\frac{\alpha - \alpha_{cr}^{0.5}}{\lambda}))$ [Fig.1], where $\alpha_{cr}^{0.5}$ - informational capacity of the network with reliability of 0.5 (both modes are equally expressed).

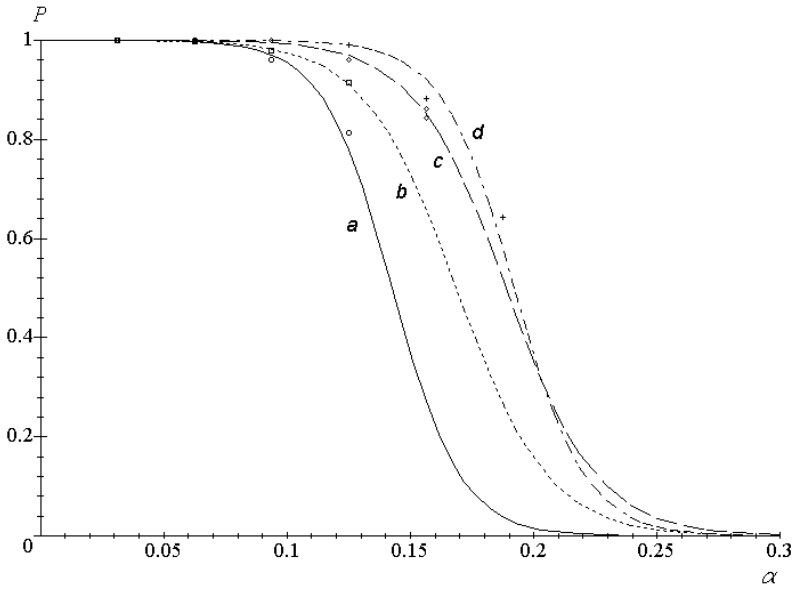

Figure 1. Approximated graph of probability *P* that final overlap is close to 1.
(a: *N*=400, b: *N*=800,c: *N*=1200, d: *N*=2200; p=0.1, S=20)

As N increases the curve of the function $P(\alpha)$ approaches step-like function. A point at which the drop occurs was taken as α_{cr}. As criterion of the drop we took the level P=0.8, i.e. $\alpha_{cr} = \alpha_{cr}^{0.8}$. As N increases values of α_{cr} also increase and saturate at some N_{sat}, the bigger S (more complex pattern), the bigger N_{sat}. For each value of S from 10 to 300 the asymptotic value of α_{cr} was determined for $N > N_{sat}$. Using (4) we obtained respective values of C_{cr}. For fixed S and p the point (C, α) moves along the beam and at $N > N_{sat}$ reaches (C_{cr}, α_{cr}). For each p such critical points form in (C, α)-plane a curve that separates phases of possibility and impossibility of factorization. As $S \to \infty$ the critical curve approaches abscise, and one may estimate the value $C_{\max}$ - maximum complexity of patterns, for which factorization is still possible. As sparseness p vanishes $C_{\max}$ increases [Fig.2].

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 387-392

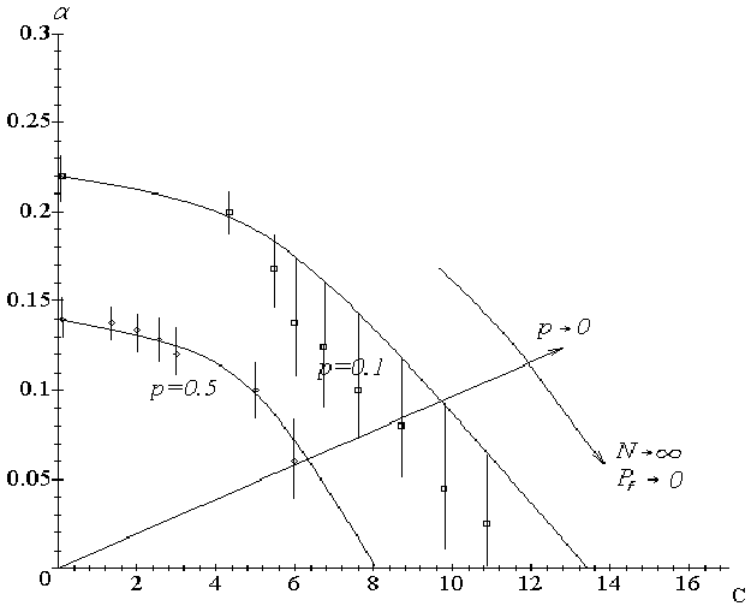

Figure 2. Critical curves in α -C plane for p=0.1 and p=0.5.
(The former is drawn taking into account that saturation by N is not reached)

Neurophysiological application

Long before occurrence of modern functional models of a hippocampus [4,5,8,12] Marr has developed conceptual model, which functional idea is close to our idea of factorization. Marr proposed that the hippocampus plays a role of the processor of the complex incoming pattern with its subsequent transfer to the neocortex as a set of "classificatory units" [10,11]. A role of temporal storage of information was assigned to a field CA3, which due to an extensive system of recurrent collaterals is a natural autoassociator. We suggest that due to presented ability of autoassociator to perform factorization, this field can carry out a function of decomposition of complex information into elementary factors.

Complex pattern could be decomposed in hippocampus into factors, which are later replayed to the neocortex in the form of coherent ensembles of neurons (Marr's classificatory units). The idea of "processing-consolidation" procedure in hippocampo-neocortical system proposed by Buzsaki [3,4] is incorporated in our model. At a stage of "learning" a complex pattern of partially processed polysensory information from neocortex modulated by theta rhythm reaches CA3, where interbound coherent ensembles (attractors) are formed. These attractors correspond to factors that encode complex pattern. At a stage of "recording" during sleep high-frequency activity in CA3 is triggered that excites those groups of target neurons in the neocortex that have strong connections with the ensembles-factors in CA3. Due to LTP-modification of synaptic connections between and within classificatory units in neocortex both prestored factors are enhanced and new ones are formed.

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 387-392

Conclusion

The proposed idea of factorization in autoassociator was partially tested. Recurrent neural network proved to be capable of extracting factors from the complex patterns structure learned using Hebbian rule. Our investigation has not yet covered the internal structure of attraction basins, number of spurious attractors and so on. This will be the aim of future research. Factorization plausibility turned out to be dependent on the absolute complexity of the patterns. It hints the idea to extend the network parameters to the real ones ($N \approx 10^5, p \approx 0.04, P\{\mathbf{J}_{ij} \neq 0\} \approx 0.02$) and use more biologically plausible neural dynamics to estimate such critical macroparameters that could be compared with their analogues in a behavioral experiment. Functional idea of factorization was shown to be applicable to neurophysiological memory models and could be further used as basis for theoretical study of memory related problems. Extensively studied spatial navigation could be an excellent test range for this idea.

References

1. Amari S, Maginu K (1988). Statistical neurodynamics of associative memory. *Neural Networks*, **1**, 63-73.
2. Amit DJ, et al (1987). Statistical mechanics of neural networks near saturation. *Annals of Physics*, **173,** 30-67.
3. Buzsaki, G (1989). A two-stage model of memory trace formation: a role for "noisy" brain states. *Neuroscience*, **31**, 551-570.
4. Buzsaki, G (1996). The hippocampo-neocortical dialogue. *Cerebral Cortex*, **6**, 81-92.
5. Eichenbaum H et al (1994). Two component functions of the hippocampal memory system. *Behav Brain Sci*, **17**, 449-518.
6. Frolov AA, Muraviev IP (1993).Informational characteristics of neural networks capable of associative learning based on Hebbian plasticity. *Network*, **4**,495-536.
7. Frolov AA, Husek D, Muraviev IP (1997). Informational capacity and recall quality in sparsely encoded Hopfield-like neural networks: analytical approaches and computer simulation. *Neural Networks*, **10** (5), 845.
8. Gluck MA (1997). Physiological models of hippocampal function in learning and memory. *Neurobiology of Learning and Memory*, **11**.
9. Hopfield JJ (1982). Neural networks and physical systems with emergent collective computational abilities. *PNAS*, **79**, 2544-2548.
10. Marr D (1970). A theory for cerebral neocortex. *Proc R Soc Lond*, B **176**, 161-234.
11. Marr D (1971). Simple memory: a theory for archicortex. *Phil Trans R Soc Lond*, B **262**, 24-81.
12. Rolls ET (1996). A theory of hippocampal function in memory. *Hippocampus*, **6**, 601-620.