Weight Modification of Recurrent Neural Network by Decorrelation

부상관성(負相關性)에 의한 순환신경망의 연결가중치 조절

Chong Ho Lee
(이 종 호)

요 약 : 순환 신경회로망의 응용에서 종종 대두되는 국지극소점을 확인하고 제거하는 효과적인 방법을 제안한다. 신경망의 학습과정에서 밝혀지는 국지극소점에 대하여 부상관성(負相關性)을 부과하여 에너지표면을 재조정 함으로서 원하는 상태에서 회로망이 안정에 도달하게한다. 이때 의사상태(擬似, spurious states)는 안정조건을 적용함으로서 확인되는데 이과정은 특별히 설계된 병렬회로에 의하여 효율적으로 처리된다. 이와같은 부학습(負學習, unlearning)의 결과로서 순환신경망의 저장용량과 수렴성능의 개선을 이룰수 있다.

Keywords: dynamic associative memory, spurious state, decorrelation

Ⅰ. Spurious states problem with associative memory

Dynamic associative memories[1] have been studied for correct restoration of noisy information or distorted images. The correlation learning rule has generally been employed for the training of such networks. Like other neural network paradigms, the recurrent neural networks are apt to be trapped at so-called spurious states. The spurious states, along with the trained states, form the basins of attractions on the energy surface during the process of learning so that the output state is stabilized at either the trained state or the spurious state whichever is closer to the initial state. The problem of such local minima has been major obstacle for reliable applications of neural networks. There have been many approaches[2,3,4,5] to reduce the chance of being trapped in such undesirably stable states. Most of the approaches try to force the output from the shallow local minima by stochastic methods or to modify the correlation matrix using Hopfield's unlearning concept[6]. Although, such methods achieve certain improvements on the chance of correct retrieval, they are subject to face criticisms: as far as there are local minima whose depths and basin sizes are unknown, the stochastic methods and the unlearning schemes and their parameter adjustment processes are not pre-determined for the unsupervised learning paradigms. And their computational complexities are immense. Moreover, some of the desirable final states may be less stable than a spurious state, in which case, a desirable output state may also be abandoned by the effort to eliminate the spurious states.

The correlation learning rule doesn't guarantee that the energy level of a trained state is lower than any of those of the spurious states, or a basin of attraction of a trained state is broader than any of those of the spurious output states.

Therefore, in order to eliminate the chance of having spurious outputs completely, the undesirable minima must be identified and removed selectively to have resculptured energy surface. This paper presents a principle way to prevent undesirable local minima which commonly degrade the performance of recurrent neural networks. This method modifies the energy surface to prohibit the undesirable minima. As was mentioned, most of the existing stochastic methods focus on overcoming the local minima which are already established by collective learning rule. The anti-Hebbian learning rule[12,13] employs the same idea of decorrelation in order to decline or compensate certain perception (in human brain or in artificial memory) which is otherwise overemphasized. However, in that case, the decorrelation is not done selectively: any strong and persistent stimuli are perceived with decay in magnitude which is interpreted as fatigue[12] in human sensibility. Such fatigues in human perception are purposive in many circumstances while they are not so in other cases. In other words, the existing schemes of anti-Hebbian (or decorrelation or unlearning) learning are subject to unspecified liability. The approach presented here is to employ the unlearning concept to only those minima which are not intended at the time of learning. This is implementable by identifying the spurious states directly from the training patterns as explained in section 3. The existence of spurious states also affects the reliable storage capacity of associative memory. The associative memory which has little or no spurious states should have higher storage capacity. So, the improvements in the radius of attractions or the probability of success retrieval should be re-examined for the suggested method. Some of the studies dealt with the improvements on the storage density as well as the correct retrieval rate for certain Hopfield-type networks[4,5].

Ⅱ. The stationary condition.

The dynamics of recurrent networks induce the outputs of the networks to converge to a certain stable state after a number of iterations when an initial input pattern is given. Such stability is acquired due to the fact that

접수일자 : 1995. 5. 16.
1차 수정 : 1995. 7. 5., 2차수정 : 1995. 8. 30.

이종호 : 인하대학교 전기공학과

the final state is one of the minima in terms of the Lyapunov's energy. As was proved by J. Hopfield[7], when the energy of an output state reaches its minimum, the state becomes immovable and stable. The condition for stability can be drawn from here. For simplicity sake, the discrete-time update scheme will be considered here.

The output at time k+1 is

$$\begin{aligned} v_i(k+1) &= sgn(net_i(k)) \\ net_i(k) &= \sum_{j=1}^{N} T_{ij} v_j(k) \qquad \text{for } i = 1,2,\ldots N \end{aligned} \tag{1}$$

where N is the number of output neurons

v_i is ith component of output bipolar binary vector v.

$$sgn(x) = \begin{cases} 1 & when\ x \geq 0 \\ -1 & when\ x < 0 \end{cases}$$

and the simplified energy E is: $E = -\frac{1}{2}\sum_i\sum_j T_{ij}\ v_i\ v_j$

At stationary state,

$$v_i(k+1) = v_i(k) \quad \text{for all i.}$$

$$\text{or, } sgn(net_i(k)) = v_i(k) \tag{2}$$

Equation (2) holds if and only if the two terms neti(k) and $v_i(k)$ are of same sign providing that the output $v_i(k)$ has reached its saturation region of the activation function. Thus, the condition for output v(k) to be stable:

$$c_i = v_i(k) \times (\sum_j T_{ij} v_j(k)) > 0 \quad \text{for all } i \tag{3}$$

In this case of using sgn(x) function as the neuron activation function, the stationary output state may not be achieved by setting $\nabla E(v) = 0$ because every output neuron may change its value end to end, that is, from -1 to 1 or 1 to -1. Thus, the output states are permissible only at the vertices of the hypercube and in this constrainded optimization problem, the energy minima appear on the boundary rather than inside the hypercube where $\nabla E(v) = 0$ is satisfied. However, the condition $v \cdot \nabla E(v) < 0$ [8] should be met for a state v not to change. This condition coincides with (3) as $v_i \cdot (\frac{\partial E}{\partial v_i}) = -\frac{1}{2} v_i(\sum_j T_{ij}\ v_j)$. The synaptic weight matrix T of the autoassociative memory is built by the following formula:

$$\boldsymbol{T} = \sum_{m=1}^{p} \boldsymbol{x}^m \cdot (\boldsymbol{x}^m)^t - p\boldsymbol{I} \tag{4}$$

where p is the number of training patterns, x, each of which is superscribed by m.

This is the same as the auto-correlation matrix obtained by Hebbian learning rule except that all of the diagonal terms, Tii, become zero in this recurrent network. If p is one, the synaptic weight obtained by (4) will do the associative retrieval correctly. However, when p is large enough, the superposition of correlation rule with respect to every non-orthogonal training pattern causes so-called crosstalk term[8,9] which may deteriorate the convergence rate or the radius of correct attraction. As the result of this, the spurious states emerge. One kind of such states occurs due to the fact that the dynamics and the energy function of the recurrent neural network have a perfect symmetry. Thus, the reverse (or complemented) states of the training patterns become stable. However, this kind of states may easily be identified and intensionally prohibited by assigning any one bit of the output vectors to known value. Other kind of spurious states is caused by the superpositional nature of the correlation learning and is a subset of the set of possible stable states which are linear combinations of odd numbers of the training patterns [9]. The actual spurious states are then determined among the combination set by the stationary conditions given in (3). From the observations above, a postulation is conceived: the spurious states may be prohibited during the training of input patterns by additional steps of compensation rather than during recall process as after-the-fact remedy.

Ⅲ. Modifying the energy surface by decorrelation

As was noted above, it is always probable that the recurrent neural networks output spurious states. The first kind of spurious states can not mislead us if a sign bit which is supposed to be positive for all the patterns is introduced. The second type of spurious states have conventionally been treated by stochastic methods. But such methods have fundamental limitation as they try to pull the output states out of the existing basins of attraction. Without question, a preventive method which compensates the false minima would be more effective. Moreover, if every spurious state is canceled, perfect retrieval with the upper-bound memory capacity would be realizable. As the correlations between the pair of bits of each training pattern are devoted to make the synaptic weight matrix and establish the storing dynamics in the recurrent neural network, the process can be reversed in order to eliminate the false minima from the energy landscape. This method of inhibitory learning is called decorrelation here. For this method, the knowledge of the spurious states in hand at the time of learning is prerequisite. This condition is met by the two theorems:

Theorem 1) The spurious states of recurrent neural network are composed of linear combinations of odd numbers of the training patterns[10].

Theorem 2) Any stationary output states of discrete-time recurrent neural network satisfy the conditions given in (3).

The decorrelation learning is to compensate the false minima which are identified by the two theorems above. For each false minimum, the weight matrix is modified by $\triangle T$:

$$\triangle T_{ij} = -\sum_{m=1}^{q} \lambda s_i^m s_j^m \tag{5}$$

where i≠j

q : number of false minima

s^m : a false minimum

λ : decorrelation factor

The complete learning process of this method is given in Fig. 1. The procedures implementing the two theorems are much less complex ways to determine the spurious states than determining the local minima by checking and comparing the energy levels of all the possible output states of n-neuron. This is especially so as p<<n. Speaking of the odd number combinations, it is observed[11] that the linear combinations of three patterns cover most of the probable stable states. That is, checking with the combinations of more than three patterns results less likely to find stable states. Thus, in practical cases, the number of admissible states to check for stable states is limited and the computational complexity of this process become tolerable, especially with the parallel hardware illustrated in Figure 2. This method should not be confused with supervised error-correction learning because during the learning process of Fig. 1, no actual recall process is exploited, and the comparison with the training patterns are not for figuring out the errors between the actual outputs and the desired outputs but for checking if the expected local minima match with the training patterns.

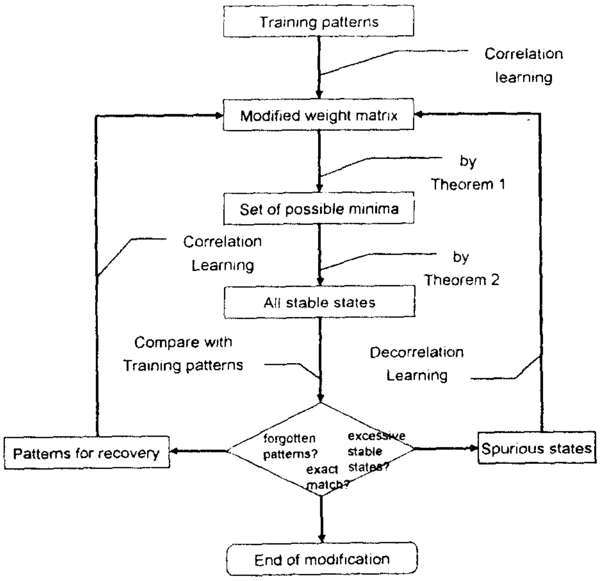

Fig. 1. Weight modification process.

Similar concepts of reversing the sign of covariance learning rule are suggested recently. Kohonen constructed a 'novelty filter'[12] which learned to be insensitive to familar features in its input to a recurrent autoassociator. The 'anti-Hebbian' synapses have also been used for lateral decorrelation of feature detectors[13,14]. A phenomena of 'unlearning' was first introduced by Crick[15] to explain the biological purpose of sleep in human and animals as a period during which unneeded information is erased and stored information is compacted. Although it has not been clearly known what active mechanism in human brain causes such reconfiguration of memory or psychological response patterns, it is generally understandable that any particular memory or passion soothes over passage of time. Soon after then, J. Hopfield presented the mathematical modelling of 'unlearning' in collective neural network by which the memory function is improved by the equalization of accessibility and the suppression of spurious memories[6].

The idea of 'unlearning' is extensively studied and practically employed here with the aids of stationary conditions and a special network to solve a set of inequalities.

IV. Numerical example

Suppose a 5-neuron, fully connected recurrent neural network to train three bipolar binary vectors, x1, x2, and x3. The correlation learning rule produce the initial weight matrix T.

$$x1 = -1\quad 1\quad 1\quad -1\quad -1$$
$$x2 = -1\quad 1\quad -1\quad -1\quad 1$$
$$x3 = 1\quad 1\quad 1\quad -1\quad 1$$

$$\boldsymbol{T}=\begin{bmatrix} 0 & -1 & 1 & 1 & 1 \\ -1 & 0 & 1 & -3 & 1 \\ 1 & 1 & 0 & -1 & -1 \\ 1 & -3 & -1 & 0 & -1 \\ 1 & 1 & -1 & -1 & 0 \end{bmatrix}$$

The linear combination of the three training vectors gives a spurious state candidate, s.

$$s = \mathrm{sgn}(\ x1 + x2 + x3\ ) = (\ -1\quad 1\quad 1\quad -1\quad 1\ )$$

As any two of the three training vectors are two Hamming distance apart in this case, s is equidistant from the three input vectors. The stabilities of four vectors are verified by the stationary criteria given in (3).

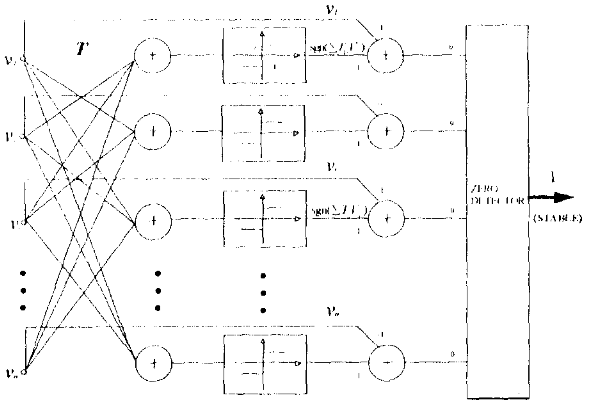

Fig. 2. Stability verification network.

The network in Fig. 2 is used to verify an output v being stable, that is, every v_i being equal to $sgn(\sum_j T_{ij} v_j)$. The zero detector outputs one when all of the inputs are zeros. This is when all the ci are positive and the output state v(k) is stable. The four stable states identified here have the same minimum energy level of -6. In order to eliminate the false minimum at s, decorrelation learning as in (5) is employed to have modified weight matrix T ' for $\lambda = 0.4$.

$$\boldsymbol{T}' = \begin{bmatrix} 0 & -0.6 & 1.4 & 0.6 & 1.4 \\ -0.6 & 0 & 0.6 & -2.6 & 0.6 \\ 1.4 & 0.6 & 0 & -0.6 & -1.4 \\ 0.6 & -2.6 & -0.6 & 0 & -0.6 \\ 1.4 & 0.6 & -1.4 & -0.6 & 0 \end{bmatrix}$$

Then, again, the stationary condition (3) is applied to verify that spurious state is not any more stable (that is, not a local minimum).

The values of c_i defined in (3) are

$$c_1^s = c_3^s = c_5^s = -1.6 < 0 .$$

Thus, it is unstable. And the energy of the spurious state has grown from -6 to -2.

$$E_s = \frac{-1}{2}\left(\sum_i \sum_j T_{ij}' s_i s_j\right) = -2$$

Fig. 3 shows how the energy of the spurious state, s, is raised by the decorrelation learning while the energies of the rest of the states being changed rather slowly. The horizontal axis is the decimal representation of all the possible 5-bit binary output vectors. The complements of the vectors are not shown because of the symmetry. If this example problem is dealt by the existing anti-Hebbian learning rule[6,12], all of the stable states, v_1, v_2, v_3 and s, are somewhat decorrelated and the energy surface becomes less rugged, and eventually, the spurious state will be removed. This is fine for the purpose of smoothing any extreme tendency or temper which is supposed to be diminished over time of sleep or rest (for the case of human). Obviously, however, our method of figuring out those unwanted attractors on the energy surface and removing them selectively in the course of learning is much effective in engineering sense.

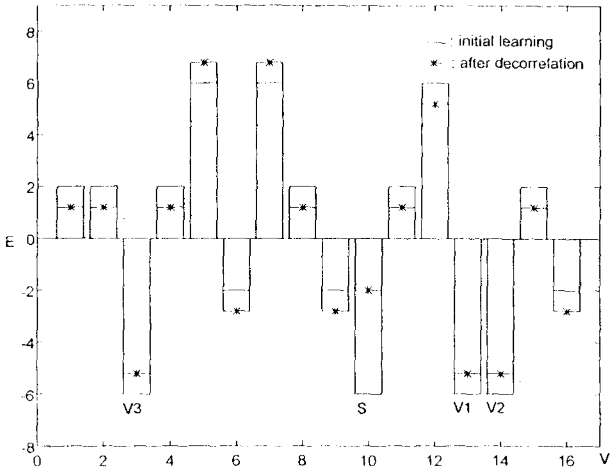

Fig. 3. Changes in energy level of output states after decorrelation learning(λ =0.4).

V. Conclusion

The local minima have been one of the serious problems associated with recurrent neural networks. Here, a principle way of preventing the undesirable minima in dynamic associative memory is presented. This method employs the idea of unlearning in order to eliminate known false minima from the energy landscape. There have been methods of reversing the learning process for various purposes. But, unless the undesirable states are known in advance, the unlearning can not be carried out properly. The method of determining the spurious states as the learning progresses is introduced in this paper and is implemented to solve the local minima problem as well as to improve the memory capacity of recurrent neural network. The verification of stability of states among the admissible set of states is done at each iteration of weight modification by a simple connectionist network. The method presented here is thus efficient because of parallelism and effective because of the acquired knowledge of spurious states. The experiments show how the energy levels of the spurious states are raised by this method so that their basins of attractions vanish.

Acknowledgement

This work was supported by Inha University Research Fund '93-'94.

References

[1] Hassoun, M., *Associative Neural Memories Theory and Implementation*, Oxford University Press, pp. 1-27, 1993.

[2] Peretto, P., "Collective Properties of Neural Networks: A Statistical Physics Approach", *Biological Cybernetics* 50, pp. 51-62, 1984.

[3] Hinton, G. & Sejnowski, T., "Optimal Perceptual Inference", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 448-453, 1983.

[4] Bachmann, C., Cooper, L., Dembo, A. & Zeitouni, O., "A relaxation model for memory with high storage density", *Proceedings of National Academy of Science*, 84, pp. 7529-7531, 1987.

[5] Potter, T., Dissertation, SUNY at Binghamton, 1987

[6] Hopfield, J. Feinstein, D. & Palmer, R., "'Unlearning' has stabilizing effect in collective memories", *Nature*, 304, pp. 158-159, 1983.

[7] Hopfield, J., "Neural Networks and Physical Systems with Emergent Collective Computational Abilities", *Proceedings of National Academy of Sciences*, 79, pp. 2554-2558, 1982.

[8] Zurada, J., "Introduction to Artificial Neural Systems", West Publishing Company, 1992. pp. 323, pp. 342.

[9] Hertz, J., Krogh, A. & Palmer, R., *Introduction the Theory of Neural Computation*, Addison-Wesley, p. 17, 1991.

[10] Amit, D. & Hanoch, G., "Spin-glass models of neural networks", *Physical Review A*, 32, pp. 1007-1018, 1985.

[11] Kanter, I. and Sompolinsky, H. "Associative recall of memory without errors", *Physical Review A*, 35 pp. 380-392, 1987.

[12] Kohonen, T. *Self-Organization and Associative Memory*, Springer-Verlag, 1989

[13] Barlow, H. & Foldiak, P., "Adaptation and decorrelation in cortex", in Miall, R. & Mitchison, G. editors, *The Computing Neuron*, Addison Wesley, Ch.4, pp. 54-72, 1989.

[14] Leen, T., "Dynamics of learning in linear feature-discovery networks". *Networks*, 2, pp. 85-105, 1991.

[15] Crick, F. & Mitchison, G., *Nature*, 304, pp. 111-114, 1983.

이종호

1953년 4월14일생, 1976년 2월 서울대학교 전기공학과졸(학사) 1978년 2월 동 대학원졸(석사) 1986년 8월 (미) 아이오와 주립대 전기및 컴퓨터 공학과졸(박사). 1979년 8월-1982년 6월 해군사관학교 전기공학과 전임강사, 1986년 8월-1989년 5월 (미) 노틀담대 전기및 컴퓨터 공학과 조교수, 1989년 8월-현재 인하대학교 전기공학과 부교수, 1994년 1월 - 1995년 1월 (미) 브라운대 두뇌및 신경망시스템 연구소 방문교수, 1991년 5월 - 1993년 5월 대한전기학회 컴퓨터및 인공지능 연구회 간사장, 주관심분야는 Computational intelligence, Neural networks,및 VLSI CAD 등임.