Acta Physica Hungarica 70 (4), pp. 317–320 (1991)

RECOGNITION AND VERIFICATION OF DISTORTED PATTERNS*

I. LOVAS

Central Research Institute for Physics
1525 Budapest, Hungary

(Received 4 June 1991)

A specific combination of Neural Networks, Expert Systems and a Correlation Analyser is studied. The system can be used for the recognition and verification of distorted patterns. The recognition of the genetic code carried by the DNA molecule is studied as an example.

There are a number of different approaches in pattern recognition research [1,2]. Here we focus our attention on a specific combination of Neural Networks, Expert Systems and a Correlation Analyser, in order to develop a method which is able to recognise and to verify distorted patterns. This same method can also be applied to reconstruct almost "forgotten" information.

The Neural Networks [3] (NN) play the role of an associative memory which is able to learn a set of patterns and later to recognise any of them even if the patterns become somewhat distorted. The Expert Systems (ES) rely on general rules which are either taught to the system or which are learnt by the system, and the output is actually a response to the input, using these rules. In the Correlation Analyser (CA) two or more inputs are searched in order to find a correlation which is assumed to exist.

If a pattern consists exclusively of random elements then it can be recognised only if one compares it element by element with a "master" pattern and proves or disproves the identity of the two patterns.

However, if we know that the pattern follows certain rules and contains some internal correlations then the situation is quite different, namely it is possible to recognise or reconstruct the pattern even if it is seriously distorted by using the extra information provided by the rules and the correlations [4].

For simplicity let us assume that the pattern under consideration can now be represented by two distinct strings of codes which carry a well known or an assumed correlation. An obvious example is a song which has a text and a tune; the text is coded by an ordered string of syllables, the tune is coded by an ordered string of notes. In the song the two sets of information are strongly correlated.

Let us analyse separately the two strings of information and at an appropriate point check whether or not the two sets show the assumed correlation.

We analyse each of the strings of codes in the following way:

*Dedicated to Prof. R. Gáspár on his 70th birthday

We teach the Neural Network all the codes which may occur in the string to be analysed. Obviously the teaching employs undistorted ("correct") codes.

The distorted code number $n + 1$ is read in by input 1.

The Neural Network tries to recall the original by randomly changing the bits of the code until a stable configuration of bits is reached. From time to time the NN is perturbed by "thermal" noise in order to prevent the system from being trapped in a shallow minimum. In the language of spin glass, such a trap corresponds to a shallow, spurious minimum of the energy functional.

The output of the NN is one of the codes previously taught to the network. The output of course may be false. The probability of a false output occurring increases with the measure of distortion of the input and decreases with the memory capacity of the network.

The output code of the NN is utilized as input into an Expert System. The ES analyses the code number $n + 1$ and checks whether it is allowed by the rules taught to the system earlier or learnt by the system up till now.

If the code number $n + 1$ is not rejected, then in the next step the ES checks to see if the code number $n + 1$ fits to the string already containing n codes.

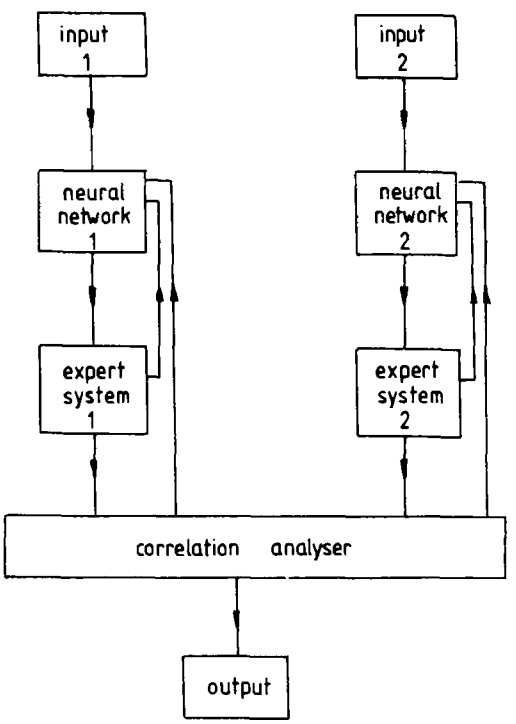

Fig. 1

If the input is rejected by the ES for either of the two reasons, the Neural Network is forced to try again to remember, hoping that it may find another min-

imum of the spin glass energy functional. Of course the number of repeated trials must be limited.

The output of the ES is a string of codes containing $n+1$ elements. This string consists of codes which on the one hand were recognised with certain reliability by the NN, and on the other hand the ES checked whether the codes correspond to the rules to be followed by the string. Now we follow the same steps working with the other string. The procedure can be generalised. If the pattern to be analysed can be or must be represented by more than two distinct strings of codes then the analysis must be repeated as many times as there are strings. Of course if it is possible then it is better to do the analysis in parallel for all of the strings.

In the next step the output signals of the Expert Systems arrive to the input points of the Correlation Analyser.

If the assumed correlation is not found then the analysis must be repeated from the NN. Of course the number of repeated trials must be limited.

The block diagram of the system is shown in Fig. 1.

An obvious generalisation is to build a network out of such units.

A simple application: the genetic code of DNA

To illustrate the essential points of the system outlined above let us look at a simple application.

This is the recognition of the genetic code of the Desoxyribo Nucleic Acid molecule. The DNA is a double helix, i.e. it can be considered as two distinct strings. These strings are connected by nucleotide pairs. DNA contains only four nucleotides: Adenine, Cytozine, Guanine and Thymine. Any of the above nucleotides may occur in arbitrary order along the strings. These are the rules to be incorporated into the Expert System.

The strings are strongly correlated for geometrical reasons. The pairs of nucleotides connecting the two strands of the double helix may only be A–C, C–A, G–T or T–G. The Correlation Analyser must give a repetition command to the NN if any other pairs are detected.

Our aim is to recognise the genetic code of a given DNA molecule. For this purpose we define the following, somewhat hypothetical "research program".

A set of biochemical measurements is performed. We assume that the result of such a set of measurements can be associated with one of N different biochemical objects.

The task of the NN is to remember all of these N possibilities and to identify the input X with one of the N possibilities even if the error of the measurements is not negligible.

The Expert System must decide whether or not X belongs to the set A, C, G, T.

Now the procedure is repeated for the other strand of the DNA molecule. The result is Y.

The Correlation Analyser must decide if the $X - Y$ pairing is allowed or not. In our "experiment" the number of possible nucleotides was $N = 16$ and they were represented by codes consisting of 256 bits. This means that the synaptic strength of the NN is represented by a 256×256 matrix.

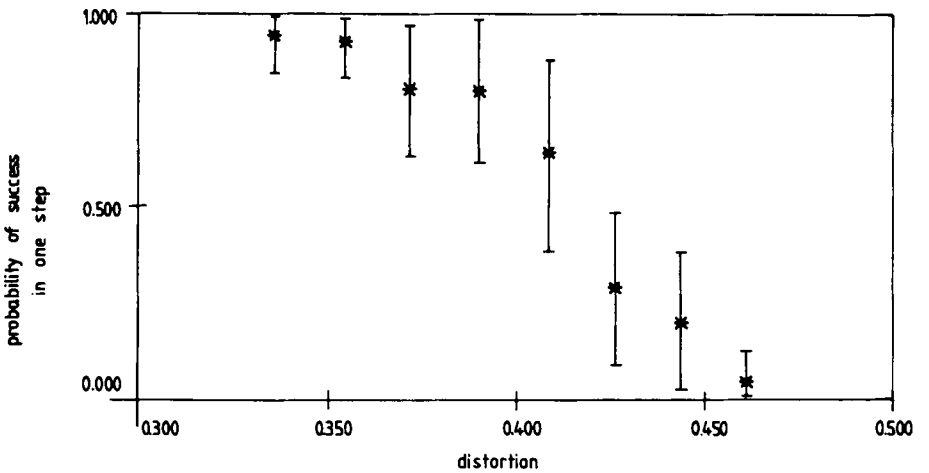

Fig. 2

During the "experiment" the degree of distortion of the input is controlled. The efficiency of the system is measured by the average velocity of the recognition and by the rate of successful identification.

The probability of successful identification S is shown in Fig. 2 as a function of the input distortion D.

References

1. D. E. Rumelhart and J. L. McCelland, Parallel Distributed Processing: Exploration in the Micro-structure of Cognition, Vol. 2, MIT Press, Cambridge, Mass, 1986.
2. T. A. Vilgis, J. Stat. Phys., *47*, 133, 1987.
3. J. J. Hopfield, Proc. Natl. Acad. Sci. USA, *79*, 2254, 1982.
4. L. Lönnbland, C. Peterson and T. Rögnvaldsson, Phys. Rev. Lett., *65*, 1321, 1990.