CONNECTIONISM IN PERSPECTIVE
R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, L. Steels (Eds.)

Problems and Trends in Integrated Neural Networks

L. Personnaz, A. Johannet, G. Dreyfus
Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris, Laboratoire d'Electronique, 10, rue Vauquelin,
75005 Paris, France

Abstract—*The paper is intended to put connectionism in perspective by comparing connectionist devices (subsequently referred to as "artificial neural networks") to classical devices, and discuss the various issues related to the silicon integration of such circuits. First the driving motivations behind the numerous attempts at integrating neural networks are outlined. The kind of artificial neurons and networks which are currently being designed, and the functions that these devices are expected to fulfill, are subsequently described. Finally, the technological issues will be outlined; we show how technologies which are used for manufacturing standard electronic devices can be used for manufacturing connectionist devices.*

Keywords—Neural network implementation, binary neurons, graded neurons, learning, feedback networks, feedforward networks, integrated circuit design, analog and digital technologies, Very Large Scale Integration, Wafer Scale Integration.

INTRODUCTION: NEURAL INTEGRATED CIRCUITS — WHY?

Integrated circuit technology is a technology whereby both the electronic components and the connections between them are manufactured on the same substrate. Manufacturing connectionist devices means a departure from the standard electronic design, since emphasis is put on the connections, not on the computing elements. Given the cost of creating a new technology, and the lack of maturity of the field of neural networks, it is clear that the designers of connectionist devices will have to use a standard electronic technology for many years: a specific technology will evolve in the future if, and only if, connectionism turns out to be very successful at solving problems. Therefore, designing and manufacturing neural networks will be a real challenge to designers for several years at least.

It is only natural to ask whether it is worthwhile to undertake such a difficult task, instead of simply performing computer simulations. Let us consider the situation as it stands now: among the various fields of applications of neural networks, pattern recognition has been the most extensively studied; as a result, there is virtually no example of a neural network performing definitely better, in terms of recognition rate, than classical methods. However, neural networks have a definite edge on other approaches as far as speed is concerned, since

they are inherently parallel computing devices. Moreover, some neural networks can be shown to exhibit, to a limited extent, fault tolerance properties. Obviously, these speed and fault tolerance characteristics will be taken advantage of if artificial neural networks are actually built. Simulations are very helpful tools for making demonstrations, validating analytical predictions, or when computational speed is unimportant; in most cases, however, real-time applications will definitely require specific chips in order to reach the desired operating speed.

Therefore, it is clear that integration is an important problem for the future of connectionist models. An adverse argument which is often expressed is a matter of schedule rather than a matter of principle: given the lack of maturity of the field, it is too early to "freeze" connectionist architectures into silicon.There is indeed a lot of truth in this: there is little doubt that the integrated neural networks that are currently in the design phase will not be those used in industrial applications, if any, within a few years. However, it is very likely that the design problems that will have to be solved for future neural networks will be very similar to the problems that are being faced now during the design of today's network chips. Therefore, the development of the final circuits will take full advantage of the knowledge which is currently being accumulated during the design of nowadays' devices, however primitive they can be.

GENERAL ISSUES: WHAT KIND OF NEURONS, WHAT KIND OF NEURAL NETWORKS?

Two kinds of neurons are usually considered : the basic McCulloch-Pitts neuron, whose state is binary (with a stepwise transfer function), and a variant thereof, whose state is graded (sigmoidal transfer function). Clearly, a binary unit is easier to design, and networks of binary units are more readily amenable to mathematical analysis, than networks of neurons with a graded response. However, the computations involved in the popular backpropagation learning algorithm requires that the transfer function of the neuron be differentiable, so that neurons with a sigmoidal transfer function are used very frequently, even in cases where the problem handled by the network is purely boolean in nature and would, in principle, require binary neurons only (for instance, classification tasks with binary input and output data). Conversely, there are indeed problems in which binary units would not be suitable (for instance, statistical classification tasks for which the state of the output units are interpreted as probabilities, signal processing tasks,...). In addition, circuits which are designed so as to be really neuromimetic use neurons with graded response, which are one step closer to neurobiology than binary neurons.

The issue of the architecture of the neural network is closely related to the choice of the neurons themselves, and to the function that is to be performed by the net. Two basic architectures are usually considered:

- feedback neural networks, which have inherently dynamical properties that make them suitable for associative memory properties, for storing and recalling either static patterns (pictures), or dynamic patterns (sequences of information, time-series);
- feedforward networks, which have interesting properties for classification purposes, and which have abilities to learn classes from examples.

It can be argued that feedforward networks are just feedback neurons in which a fraction of the synaptic coefficients are forced to zero. Therefore, a fully connected network can be used, in principle, for implementing any network architecture; for this reason, most current chip designs aim at building fully connected binary neurons.

From a technical point of view, the silicon implementation of binary neurons is relatively straightforward, whereas the design of graded neurons is much more cumbersome: it requires large silicon areas if digital technology is used; analog techniques allow a much easier implementation of sigmoidal transfer functions, but they have other problems (the issue of digital vs. analog techniques will be discussed below). The difficulty of building non-binary neurons greatly hampers the design of feedforward networks. This situation will probably change soon: recent work shows that multilayer feedforward networks of binary neurons can be trained without using the backpropagation algorithm.

FUNCTIONAL ISSUES

A key element in the design of a neural network circuit is the task(s) that it will have to fulfill. The alternative is the following :

- General purpose network or *ad hoc* network?
- Integrated, on-chip learning, or off-line learning?

Most reported designs implement general-purpose networks with off-line learning. The reason for that is probably the fact that they are basically research prototypes; they are designed in order to test architecture ideas, and to get a quantitative assessment of the technological problems and limitations that arise. There are two notable exceptions to the above statement: the neuromimetic chips such as C. Mead's retina (Mead et al., 1988), which are specifically designed to mimic some definite pieces of "wetware", and the chips designed for character recognition (Graf & Hubbard, 1989).

The issue of on-line vs off-line learning will be important in the future. Since speed is one of the main assets of neural networks, it is natural to try to take advantage of this characteristics both during the training phase and during the classification or retrieval phase. Therefore, attempts are being made at designing networks which are able to modify their own synaptic coefficients; this can be done effectively inasmuch as the hardware required to implement the learning rule is basically the same as the hardware needed to implement the neurons themselves (Personnaz et al., 1989).

TECHNOLOGICAL ISSUES

In this section, we shall discuss the various options that can be taken for designing and manufacturing artificial neural networks. The first issue, which is probably the most important, is the choice between analog and digital techniques. The second issue is related to the scale of integration required for building the networks. Finally, less classical technologies will be discussed.

1) Analog *vs.* digital:

An individual neuron is an object which performs a weighted sum that is subsequently subjected to a transfer function; the latter can be either a step function or a sigmoidal function. At first sight, analog technology seems ideally suited since an operational amplifier and a couple of additional components can perform these functions in a very cost- and area-effective way. Analog devices, however, have their own drawbacks which led to a drastic decrease in the use of that technology during recent years; it is sometimes argued that, within a few years, analog technology will be used only in analog-to-digital and digital-to-analog converters, and in switched-capacitor filters. The main problem is the accuracy required for the synaptic weights in order for the network to perform efficiently. This is highly dependent on the application for which the network is built; when the network is intended to perform a classification task, this is a very crucial problem. It has been shown that neural networks used as associative memories for storing and retrieving uncorrelated prototypes require a very low accuracy, and that binary (encoded on a single bit) or ternary synapses yield satisfactory results. In the general case, the number of bits required to code for a synaptic weight is O(log2n), where n is the number of binary neurons. In analog technology, synapses are implemented as resistors or transistors, which have an inherently low accuracy and low noise immunity. For neurons with sigmoidal transfer functions, an additional problem arises, namely, the accuracy required for the computation of the sigmoid; this is a very difficult problem, which, to the best of our knowledge, has not been studied by analytical means.

To summarize, analog technology is perfectly suited to applications of neural networks which

i) do not require a high accuracy for the synaptic weights and the transfer function,

ii) are relatively immune to noise, or even require some noise, as the Boltzmann machine does (Alspector et al., 1988).

In the next section, we shall discuss two implementations using analog techniques.

Conversely, digital techniques enjoy high noise immunity, and can represent synaptic weights with an arbitrary precision; moreover, the values of the latter can be changed easily, whereas the values of resistors or the resistivity of transistor channels cannot be altered accurately in any straightforward fashion. In addition, the dynamics of the neurons can be chosen freely, whereas analog neurons cannot be clocked very easily. However, these advantages are obtained at the expense of an increase in silicon area. At the present time, the implementation of approximately one hundred neurons in standard digital technologies leads to a rather bulky chip; note, however, that the design of such a chip is not a intractable problem because it is basically cellular in nature, in contradistinction to the design of, e.g. a microprocessor or an application-specific integrated circuit, which may require a large amount of "random" logic.

2) Very Large Scale Integration (VLSI) *vs* Wafer Scale Integration (WSI)

At the present time, VLSI technology has led to the design and manufacturing of commercial circuits with a maximum area of approximately 1 cm2 and a maximum number of transistors on the order of 1 million. An alternate possibility might be the use of the whole area of the

silicon wafer (with a diameter on the order of 10 cm) to build a single, huge circuit. The problems encountered in the latter technology are twofold:

i- the complexity of the design of such a circuit is extremely high, unless the circuit is cellular in nature, which is the case of neural networks;

ii- the yield problem is crucial; at the end of the fabrication process, a (possibly large) percentage of the modules present on the wafer do not operate satisfactorily. Therefore, all the circuits must be tested, and those which are operating satisfactorily must be connected together, during the so-called "reconfiguration" phase, in order to produce a working device; this technique implies obviously that the design must be highly redundant, identical replicas of the same circuit being present in several locations on the wafer, in order to be reasonably sure that at least one of them will operate properly. Therefore, the inherent fault tolerance that neural networks are claimed to exhibit make them very attractive candidates for wafer-scale integration, since the reconfiguration phase might be absent, or at least much simpler than for conventional systems. Investigations along these lines are underway.

3) Silicon technology *vs* others

All the previous sections dealt with various aspects of silicon technology, its capabilities and its limitations. However, alternate technologies might prove useful, in conjunction with silicon. Opto-electronics, for instance, is an attractive technology because it might help solving the interconnection problem. It has been used for neural network simulators, and might become attractive if it could be amenable to integration, which is not the case at the present time. The problem of interconnections could be alleviated to a considerable extent if three-dimensional connections were possible; in this respect, MOS transistors made of inorganic molecular semiconductors (doped phtalocyanines) are very attractive candidates because they are manufactured in thin film technology at room temperature, and can be stacked; this technology is far from maturity, but is definitely worth considering at the laboratory level.

REPRESENTATIVE EXAMPLES

The present section is by no means intended to give an exhaustive overview of the state of the art in neural circuit design. We shall describe circuits which are representative of the various issues discussed above.

As mentioned previously, two types of integrated circuits have been built:

i) dedicated networks, usually in analog technology,

ii) general-purpose networks, such as feedback nets, multilayer feedforward nets, or Boltzmann machines.

Typical of the first kind are the machines described in Mead et al. (1988) and Graf & Hubbard (1989). In both cases, the circuits are dedicated to a very specific task, and do not require any training. However, the generic problems that arise in the design of neural circuits are present: cellular structure, large number of interconnections, computation of weighted sums; they all exhibit a high degree of parallelism. C. Mead's retina is basically an array of operational amplifiers and of resistors, arranged in such a way as to detect the motion of

objects. It is designed in a neuromimetic fashion, and indeed the response of the artificial retina is very similar to that of a natural retina, as far as motion detection is concerned.

The networks described in Graf and Hubbard (1989) are dedicated to preprocessing tasks in a handwritten character recognition device; they use binary synapses, the strengths of which are the components of the prototype vectors; an input binary vector is thus compared in parallel to all the prototype vectors, the comparison consisting either in computing the number of zeroes common to the unknown vector and each prototype vector, or the total number of common bits, depending on the specific application. Each synapse consists of MOS transistors used as resistors or programmable switches. The summations are performed by operational amplifiers. In the latest design of the authors, 46 vectors of 96 binary components each are stored, and the comparison of an unknown vector to all the stored vectors is performed in 100 ns. When such speeds are reached, input/output operations tend to become a bottleneck.

We now turn to the design of fully digital circuits. Both described devices implement fully interconnected feedback networks. The circuit described in Murray et al. (1988) is characterized by the fact that the state of each neuron can take on five values. The design of Personnaz et al. (1989) implements binary neurons with on-chip learning.

The network designed by Murray et al. (1988) is basically a rectangular grid of synapse-processors, each neuron being associated to a column of processors. The computation of the potential of the neurons being performed simultaneously through the columns, the need for long-range wiring is reduced. However, the partial sums are transmitted serially from one "synapse" to the next on several bits (16 bits if 8-bit synaptic coefficients are used). The state of each neuron is encoded on two bits plus a sign bit, thereby allowing five possibilities: 0, ± 0.5, ± 1.Therefore, no multiplicator is required, the operations being only a shift or a boolean operation on the sign bit. Thus, each synapse processor includes a shift register and an adder-subtractor. Once the potential of each neuron has been computed, the updated state is obtained by a thresholding operation according to the 5-state transfer function.

In the network described in Personnaz et al. (1989), an alternate approach is taken, whereby neuron-processors are used instead of synapse-processors. Each neuron is associated to a memory containing the relevant synaptic coefficients; it performs the product of the binary state of a given neuron with the corresponding coefficient (which is simply a boolean operation on the sign bits), adds it to its potential, and subsequently passes the state to the next neuron. Therefore, all n neurons perform the computation of their potential in n steps. The only information that are passed between neurons are the states, encoded on a single bit. In addition, the same architecture can be used with very little additional hardware for performing the training of the network by Widrow-Hoff's rule.

Additional references to circuit designs are: Blayo & Hurat (1989), Duranton et al. (1989), Faure & Mazaré (1989), Moopenn et al. (1988), Sivlotti et al. (1985), Tsividis & Anastassiou (1989), Verleysen et al. (1989).

CONCLUSIONS

We have attempted to give an overview of the problems and trends in the design of silicon-integrated connectionist devices. There is little doubt that this kind of research will become increasingly popular in the next years, especially if neural networks find large-scale

applications in various fields. Although the experimental chips that are in existence at the present time will probably not be in commercial use, the design style that they induce, the interconnection and test problems that they generate, will certainly be beneficial to the community of chip designers. Thus, when the time of commercial applications comes, the silicon implementation of the required neural networks will be at hand.

REFERENCES

Alspector, J., Allen, R.B., Hu, V., Satyanarayana, S. (1988). Stochastic learning networks and their electronic implementation. In: D.Z. Anderson (ed). *Neural Information Processing Systems, Natural and Synthetic.* American Institute of Physics.

Blayo, F. & Hurat, P. (1989). A systolic architecture dedicated to neural networks. In: L. Personnaz & G. Dreyfus (eds). *Neural Networks from Models to Applications.* I.D.S.E.T., Paris.

Duranton, M., Gobert, J. & Mauduit, N. (1989). A digital VLSI module for neural networks. In: L. Personnaz & G. Dreyfus (eds). *Neural Networks from Models to Applications.* I.D.S.E.T., Paris.

Faure, B. & Mazaré, G. (1989). A VLSI asynchronous cellular architecture dedicated to multilayered neural networks. In: L. Personnaz & G. Dreyfus (eds). *Neural Networks from Models to Applications.* I.D.S.E.T., Paris.

Graf, H.P. & Hubbard, W. (1989).VLSI neural network for fast pattern matching. In: L. Personnaz & G. Dreyfus (eds). *Neural Networks from Models to Applications.* I.D.S.E.T., Paris.

Mead, C. & Mahowald, M.A. (1988). A silicon model of early visual processing. *Neural Networks,* **1**, 91-97.

Moopenn, A., Langenbacher, H., Thakoor, A. P. & Khanna, S. K. (1988). A programmable binary synaptic matrix chip for electronic neural networks. In: D.Z. Anderson (ed). *Neural Information Processing Systems, Natural and Synthetic.* American Institute of Physics.

Murray, A.F., Smith, V.W., & Butler, Z.F. (1988). Bit-serial neural networks. In: D.Z. Anderson (ed). *Neural Information Processing Systems, Natural and Synthetic.* American Institute of Physics.

Personnaz, L., Johannet, A., Dreyfus, G., & Weinfeld, M. (1989). Towards a neural network chip: a Performance Assessment and a Simple Example. In: L. Personnaz & G. Dreyfus (eds). *Neural Networks from Models to Applications.* I.D.S.E.T., Paris.

Sivilotti, M., Emerling, M. R., & Mead, C. (1985). A novel associative memory implemented using collective computation. In: *Proc. Chapel Hill Conf. on VLSI.*

Tsividis, Y.P., Anastassiou, D. (1987). Switched-capacitor neural network. *Electronics Letters* **23**, 958-959.

Verleysen, M., Sirletti, B., Jespers, P. (1989). A new VLSI architecture for neural associative memories. In: L. Personnaz & G. Dreyfus (eds). *Neural Networks from Models to Applications.* I.D.S.E.T., Paris.