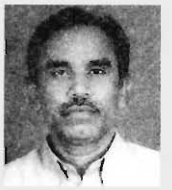
Optoelectronic Implementation of Neural Networks

Use of Optics in Computing

R Ramachandran

This article discusses the optical architectures and techniques in implementing neural networks as they apply to optical pattern recognition and optical information processing systems.



R Ramachandran is in the faculty of the Department of Electronics and Communications at Sri Venkateswara College of Engineering, Sriperumbudur. He is doing research towards his doctoral degree in optical computing techniques at Anna University, Chennai.

Introduction

A conventional computer is adept at mechanically executing the instructions in an algorithm. It can solve some classes of computational problems thousands of times faster and more accurately than humans. But the computer cannot match the memorization and recollection capability of the human brain which regularly and effortlessly solves pattern recognition problems. Humans can recognize a man in a crowd from a mere glimpse of his face. The biological neural system architecture is completely different from the computer and this difference in the structure leads to the unique capability of the brain to solve a variety of ill-structured problems. The human brain processes information in a coordinated manner with the aid of a network of a very large number of densely interconnected, relatively simple decision-making elements, the neurons. It is estimated that the brain has of the order of 10^{10} to 10^{11} neurons, each making 10^3 to 10^4 connections to neighbouring neurons. The total number of synaptic interconnections where information is stored is extremely large, approaching 10^{14} or 10^{15}. The number of connections in a neural network is in sharp contrast to the very sparse connectivity that exists in electronic circuits of computers. The difference in capabilities stems directly from the collective nature of information processing in the brain as opposed to the predominately segmented and largely

It is estimated that the brain has of the order of 10^{10} to 10^{11} neurons, each making 10^3 to 10^4 connections to neighbouring neurons.

Devices designed to model the workings of the human brain by emulating its anatomic structure are called artificial neural networks (ANN).

serial processing architectures of conventional digital computers. The difference between the way a neural net and a digital computer approach problems is evident while solving simple arithmetic problems. When we are asked what 3 times 8 equals, we do not need to calculate the result, we immediately give the answer 24 because in early age we memorized multiplication tables; we simply recognize the prompt and recall the associated answer.

Artificial Neural Network

Devices designed to model the workings of the human brain by emulating its anatomic structure are called artificial neural networks (ANN). Like the brain, they would consist of a large number of simple processors that are densely interconnected. An objective of an ANN is to stimulate functions of biological neural networks, such as learning, adapting and copying by means of parallel dispersive processing based on neural network models. In other words, ANN research seeks to simulate the structure of the massively interconnected biological neural network by which information processing can be performed associatively and in parallel (see *Resonance*, Vol.1, No.2, pp.47–54).

Figure 1 shows how neurons are assembled into interconnected layers. Each interconnection in an ANN has a weighted strength,

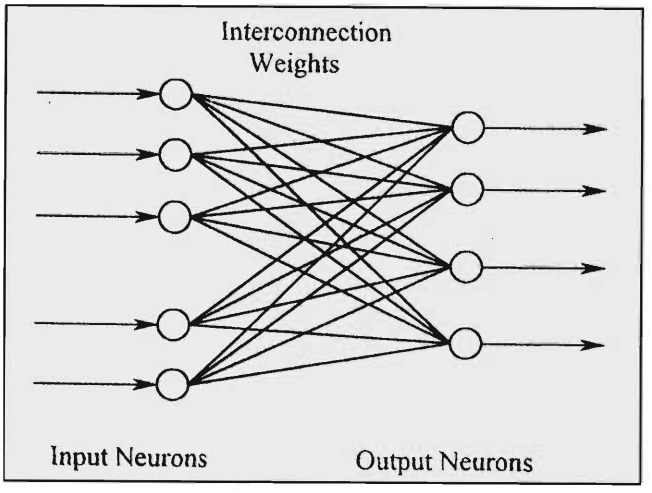


Figure 1. Single-layer neural network.

Box 1. Applications of Artificral Neural Networks

The main attributes of neural processing are its nonlinear and adaptive learning capability, which enables machines to recognize possible variations of the same object or pattern and/or to identify unknown functions and mappings based on a finite set of training data, which can be noisy with missing information. Growing members of neural- network solutions to real-world signal-processing problems have been reported. Some recent examples are: Optical character recognition, pattern classification, automobile control, aircraft control, speech synthesizing ,neural vision system, real time face recognition representing 3-D objects and image processing.

Apple computer's Newton message pad is a neural network based character classifier which is also used to provide robust recognition of hand-printed English text. Engine idle and misfiring controls are being developed by Ford Corporation using recurrent neural network. Neural networks have been shown to be applicable in the separation of blindly mixed signals, where neither knowledge of the signals nor of the mixing process is available. The neural vision system is used for a number of complex real-world image processing applications such as computer assisted diagnosis of breast cancers; biomedical vision to analyze neurobiological disorders.

much like a programmable circuit array or valve that attenuates or magnifies transferred signals. Each neuron is connected to every neuron in the preceding layer and each neuron performs a weighted sum of the input signals and nonlinear transformations and broadcasts the output to the neurons in the next layer. The transfer function of the network can be a threshold function, i.e. a step function, making the state of a neuron binary (0 or 1), or a sigmoidal function, which gives a neuron analogue values. The weights can be changed by a learning law, a rule that modifies the weights in response to the input signals and the value supplied by the transfer function. The learning law allows the neuron's response to change with time, depending upon the nature of the input signals. This is the means by which the network adapts itself to the environment and so organizes information within itself; in short, it learns. There are generally two kinds of learning, supervised and unsupervised. Supervised learning requires a teacher to supply the network with both input data and desired output data as training examples : in other words, references. The network must be taught when to

For a single layer ANN, in general, the number of interconnections is the square of the number of neurons comprising a given network.

learn and when to process information, but it cannot do both at the same time. In unsupervised learning, the network is given input data but no desired output data; instead, after each trial or series of trials, it is given an evaluation rule that evaluates its performance. It can learn an unknown input during the process. It mimics the human's self learning ability. Many ANN models have been developed in the past few decades. To name a few, Hopfield, perceptron, error-driven back propagation, and Boltzmann machine are the supervised learning models; the adaptive resonance theory (ART), neocognitron, Madaline, and Kohonen self-organizing map models are among the best known unsupervised learning models.

Electronics or optics may be used to implement the ANN. The advantages of electronic implementation derive from the fact that it is based on a very mature technology. The neurons (simulated by transistors) and the interconnections (implemented with wires, resistors and transistors) are both fabricated on the same planar surface. Typically, the area needed for interconnections is a large fraction of the available area and this severely limits the size of the network that can fit on a single chip. In electronic implementation the wires cannot cross each other and are to be kept separated by atleast a minimum critical distance. Spurious coupling between them makes it difficult to implement massive interconnections. For a single layer ANN, in general, the number of interconnections is the square of the number of neurons comprising a given network. For example, a fully interconnected network of 10^4 neurons requires 10^8 interconnections, which is beyond the reach of the state of the art VLSI technology. This practical restriction places an effective limit on the number of wires that can be placed on a chip and hence on the amount of data communication that can take place on the chip. Also on chip, communication delays increase linearly with interconnection lengths and interchip communications suffer from the large off-chip parasitic capacitances which dramatically increase interconnection delay and power consumption.

Optical Neural Networks

Restrictions on the number of interconnections can be alleviated by the promising technology 'optics'. Optics offers advantages in realizing the parallelism and massive interconnectivity required in fabricating ANN. In the optical implementation of neural network, the processing elements are provided with optical input/output capability and the third dimension, normal to the processing plane, is used for global interconnections. Light beams unlike current carrying wires can pass through a three dimensional space without interfering with each other. Because optical processing elements communicate through beams of light, they can be interconnected with one another without attaching wires between pairs of elements.They need not be confined to the restrictive planar configurations of silicon chips. Indeed, optical interconnections are being considered as a means of relieving communication bottlenecks encountered in VLSI chips. Owing to the high speed of light propagation, optical interconnection delay is practically independent of the interconnection length. In a hybrid optoelectronic ANN system the processing units are electronic, but the connections between them are optical, typically consisting of light sources and light detectors fabricated on the same chip as the processing units.

Because optical processing elements communicate through beams of light, they can be interconnected with one another without attaching wires between pairs of elements.

Matrix Vector Multipliers

The majority of optical neural networks fall into one of the two categories: Matrix vector multipliers and holographic correlators. The hardware implementation of the simple model requires only two operations to be performed. A matrix product of the vector of input activities u with the matrix of synaptic weights T,

$$u'_i = \Sigma_j T_{ij} * u_j ,$$

where u'_i denotes the sum of all signals that the i^{th} neuron in the output layer receives. A thresholding operation is then applied giving

$$v_i = s\,(u'_i),$$

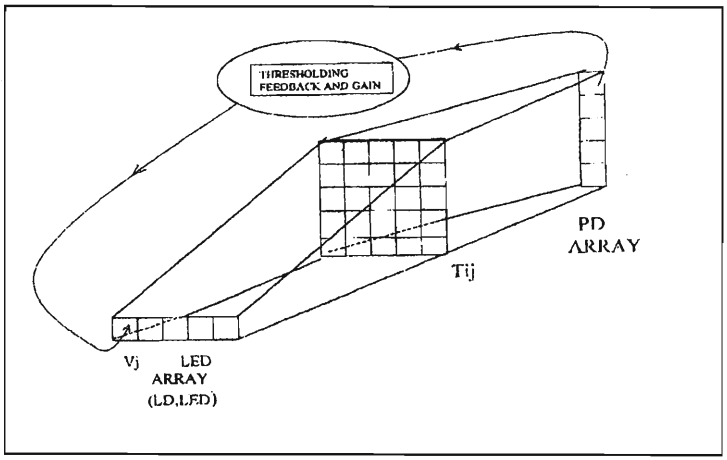


Figure 2. Optoelectronic implementation of neural network.

where v_i denotes the activity of i^{th} neuron in the output layer, and s represents the bounded activation function of a neuron. It is typically monotonically increasing, such as the sigmoidal function

$$\sigma(x) = 1/(1+e^{-x}).$$

Optoelectronic implementations of Hopfield model ANN basing vector-matrix multiplier uses an array of light emitting diodes (LEDs) to represent the logic elements or neurons of the network. Their state (on or off) can represent unipolar binary vectors such as the state vectors stored in the memory matrix T_{ij}. Global interconnection of the elements is realized as shown in *Figure 2* through the addition of nonlinear feedback (thresholding, gain and feedback) to a conventional optical vector matrix multiplier in which the array of LEDs represents the input vector, and an array of photodiodes (PD) is used to detect the output vector. The output is thresholded and fedback in parallel to drive the corresponding elements of the LED array. Multiplication of the input vector by the T_{ij} matrix is achieved by horizontal imaging and vertical smearing of the input vector that is displayed by the LEDs on the plane of the T_{ij} mask (by means of an anamorphic lens system omitted from the figure for simplicity). A second anamorphic lens system (also not shown) is used to collect the light emerging from each row of the T_{ij} mask on individual photosites of the PD array. In this system feedback is achieved by electronic wiring. It is possible and preferable to dispose of

Optoelectronic implementations of Hopfield model ANN basing vector-matrix multiplier uses an array of light emitting diodes (LEDs) to represent the logic elements or neurons of the network.

Box 2. Current Status of Optical Neural Networks

The first optical implementation of neural networks was proposed by D Psaltis . Two dimensional programmable optical neural network was implemented by T Lu and others. Pattern translation has been accomplished using the optical hetero association neural network model by F T S Yu and T Lu, L Zang and others implemented a second – order single layer neural network using only one input spatial light modulator and no intermediate electronic processing. D Psaltis and others implemented an adaptive optical neural network using photo refractive crystals and realized interconnection density of 10^8 to 10^{10} per cm^3 . B Javidi and others designed a correlator based two-layer neural network associated with a supervised perceptron learning algorithm for real-time face recognition .

electronic wiring altogether and replace it by optical feedback. This can be achieved by combining the PD and LED arrays in a simple compact hybrid or monolithic structure that can also be made to contain all ICs for thresholding, amplification, and driving of LEDs. Bulky free space optics of the optical vector-matrix multiplier can be replaced with a sandwich structure consisting of an LED array with line shaped LEDs, an optical weight mask, and a PD array with lineshaped photodiodes. Considerable versatility can be added to such a chip by employing a computer-controlled dynamic nonvolatile spatial light modulator as a programmable synaptic mask (memory mask).

Holographic Neural Networks

Holography represents a link between optics and ANN. It was recognized early that when a hologram of an object is recorded with a coded reference, the object is reconstructed when the hologram is illuminated with the same reference beam. Moreover, the reference beam can be reconstructed by illuminating the hologram with the object beam. We can therefore think of the reference and object beams as being associated with each other. The distinction in this case between an object and a reference becomes unimportant; we simply think of two patterns being associated with each other. Distorted partial versions of either pattern can also reconstruct the other, and if a portion of the recorded hologram is eliminated, the recorded pattern can be reconstructed with fidelity that degrades gradually as the portion

Holography represents a link between optics and ANN.

Holograms are used to connect each neuron with others in the same or adjacent processing planes.

of the hologram that is removed increases. These properties are reminiscent of the behavior exhibited by the outer product associative memory model Hopfield neural network. The most promising method of optical interconnections involves the use of holography. Though holograms are best known for their ability to generate three dimensional images, they are more generally capable of redirecting light in a programmable fashion. Holograms are used to connect each neuron with others in the same or adjacent processing planes. In contrast with implementations, in which the specifications of the connection patterns must be stored separately from the connections themselves, holographic media can simultaneously provide both the massive physical interconnectivity and the large memory required to specify the connections. This duality is particularly useful in adaptive networks.

Figure 3 shows the holographic analogue of two neurons. The output of each neuron is a light beam, and the activity of the neuron is coded in the amplitude or intensity of the optical signal. The input of each neuron is a light detector which senses the amount of light that is directed towards it. A holographic grating is placed in the path of the output beam of neuron B, which diffracts the incident light. The direction of the diffracted beam is determined by the period and the orientation of the grating. With an appropriate holographic grating, light from neuron B illuminates the detector of neuron A. In this way a signal is generated in A as a result of the activity in B and we say that the hologram connects the two neurons. The strength of the connection can be modified by adjusting the modulation strength of the holographic grating. Here the output light beam plays the

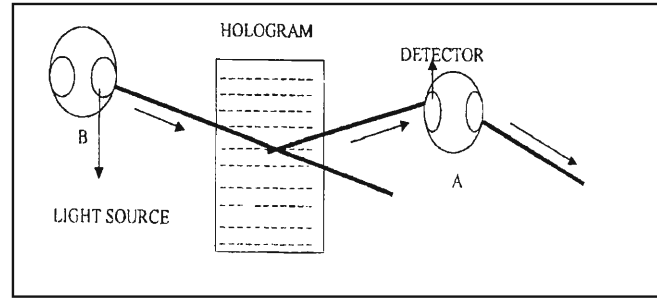


Figure 3. Holographic analogue of two neurons.

role of the axon, broadcasting the signal from each neuron. The holographic grating plays the part of the synapse directing the signal from one neuron to the next, and the optical pathways along which light is transferred from the hologram to the detector area of the neuron are analogous to the dendrites. The device consisting of the optical beam generator, the detector and the circuits that process the detected signal is reminiscent of the soma of the neuron.

Many such neurons can be connected together to form an optical neural network. Suppose that the third neuron is to be connected to neuron A as well. This can be optically simulated by superimposing a second holographic grating in the same crystal that was used to store the interconnection pattern between the first two neurons. This second grating has a distinct spatial orientation and period and it is tuned to redirect light from the third neuron onto the same input (the detector) of neuron A. Here, there is a difference between a real neuron and its holographic realization: the synapses are not localized at the intersection between axon and dendrite, but are implemented instead in a distributed manner, each synapse sharing the entire volume of the holographic medium. Also, secondly, in real neurons, some processing tasks can take place on the branches of the dendritic tree, but in the optical simulation all integration and computation tasks are concentrated on the integrated unit, 'neuron'. The advantages of the distributed holographic synapses are high storage density and ease of fabrication. The number of distinct synapses that can be packed in a hologram of a volume v is v/λ^3, where λ is the wavelength of the light. This corresponds to 10^{12} synapses per cm^3 for the typical operating wavelength $\lambda \sim 1\mu$m. The distributed holographic storage of the weights is also the ultimate in simplicity in terms of device fabrication: it involves only crystal growth. Fixed holograms can be used in applications where learning is accomplished prior to actual use. Often, however the network must continuously adapt itself to changing conditions for better overall performance. Dynamic holographic media, such as photorefractive crystals allows

In real neurons, some processing tasks can take place on the branches of the dendritic tree, but in the optical simulation all integration and computation tasks are concentrated on the integrated unit, 'neuron'

The simplest form of holographic associative memory is realized by recording the interference between associated image pairs.

continuous modification of interconnection patterns. In photorefractives, the interference between desired input and output light patterns moves charges in the crystal, creating an electric field distribution which modulates the index of refraction. This varying index acts as a holographic grating that reproduces a desired output in the presence of its associated input. The portion of light emanating from one neuron, which is redirected as light converging to another, is determined by parameters in the writing process. By controlling these parameters, the efficiency of the coupling between each pair of neurons may be specified.

The simplest form of holographic associative memory is realized by recording the interference between associated image pairs. This is equivalent to recording the hologram of an image using its associated pattern as the reference beam. The interference patterns between two images placed at the input and training planes are holographically stored in the photorefractive crystal. Presentation of either of the two images at the input plane reconstructs its associated image as the output.

Figure 4 shows a particular realization of a holographic neural network architecture. Each resolution element (or pixel) at the first neural plane can be a location for a neuron. The light from a pixel is collimated and diffracted by a holographic grating. The diffracted light is focussed by a lens onto a pixel at the output neural plane. The portion of the light that is not diffracted by the hologram is reflected by a phase conjugating mirror (PCM). A PCM unlike a conventional mirror, reflects a collimated beam back along its original path. A PCM is a nonlinear crystal that forms a hologram of the incident beam. The recorded hologram is then illuminated from the opposite direction, resulting in a

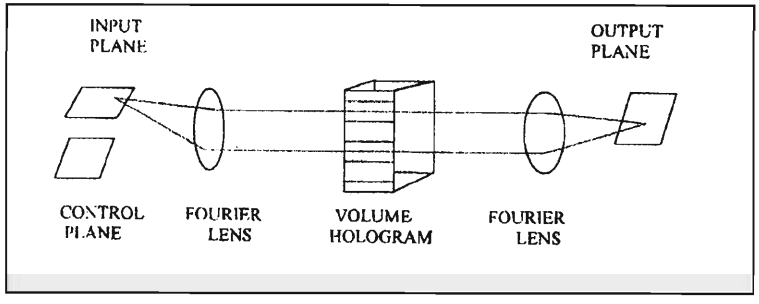


Figure 4. Optical holographic hetero-associative memory.

reconstructed beam that precisely retraces the path of the beam incident on the PCM. Light from the output neuron propagates towards the left in the figure and is collimated by the lens. This beam and the one that is reflected by the PCM form a sinusoidal interference pattern at the hologram. The interference pattern exposes the hologram and a sinusoidal grating is recorded. This arrangement ensures that the grating will interconnect the input and output pixels or neurons that were used to record it. The strength of the connection is controlled by varying the exposure during recording. The connection is stronger when the light is brighter or the exposure time longer. If x_i and x_j are the amplitudes of the light from the input and output neurons, respectively, then the change in the strength or the weight of the connection formed by the grating is

$$\Delta W_{ij} = a\, x_i\, x_j ,$$

where x_i and x_j can be, in general, bipolar signals. When the sign of x_i matches the sign of x_j then W_{ij} increases; otherwise it decreases. If each of the input plane and output plane contains $N * N$ pixels or neurons then we need $N^2 * N^2$ interconnections, which implies that N^4 weights or gratings must be stored in the hologram. The optical setup given above is a typical example for an associative memory neural network model. Photorefractive crystal has been used as the holographic medium. Holographic connections that will associate multiple input and output image pairs are established by superimposing a sequence of holograms.

Conclusion

The optical neural networks described in this article establish a link between neural network modeling and optics. We hope this will prove useful in the implementation of optical information processing systems. A basic compatibility exists between what optics has to offer and what is required for the simulation of neural network models. As new models emerge and their sophistication increases, we can expect that optical implementations of these models will continue to show advantages over other approaches.

Suggested Reading

[1] Simon Haykin. *Neural Networks*, Maxwell Macmillan International, Singapore.

[2] James A Freeman. *Neural Networks Algorithms, Applications and Programming Techniques*. Addison – Wesley. Publishing company.

[3] T S Yu Francis. *Optical storage and Retrieval Memory Neural Networks and Fractals*. Marcel Dekker Inc, NewYork.

[4] Dror G Feitelson. *Optical Computing*. The MIT Press Cambridge, London.

[5] M H Hassoun. *Fundamentals of Artificial Neural Networks*. Prentice Hall of India, New Delhi, 1998.

[6] J A Anderson. *An Introduction to Neural Networks*. Prentice Hall of India. New Delhi, 1998.

Address for Correspondence
R Ramachandran
Department of Electronics & Communication Engineering
Sri Venkateswara College of Engineering
Sriperumbudur 602 105