Concept of a neural system for real-time evaluation of spectroscopic measurements

Redei, L., Fried, Miklos, Lohner, Tivadar, Polgar, O., Barsony, Istvan, et al.

L. Redei, Miklos Fried, Tivadar Lohner, O. Polgar, Istvan Barsony, Sz. Forizs, Zoltan G. Horvath, Josef Humlicek, H. Wallinga, "Concept of a neural system for real-time evaluation of spectroscopic measurements," Proc. SPIE 3573, OPTIKA '98: 5th Congress on Modern Optics, (19 August 1998); doi: 10.1117/12.320999

Event: OPTIKA '98: Fifth Congress on Modern Optics, 1998, Budapest, Hungary

CONCEPT OF A NEURAL SYSTEM FOR REAL-TIME EVALUATION OF SPECTROSCOPIC MEASUREMENTS

L. Rédei, M. Fried, T. Lohner, O. Polgár and I. Bársony; Research Institute for Technical Physics and Materials Science, H-1525 Budapest P.O.Box 49, Hungary
Sz. Fórizs, Z. G. Horváth; Research Institute for Solid State Physics, H-1525 Budapest P.O.Box 49, Hungary
J. Humlicek; Department of Solid State Physics, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic
H. Wallinga; MESA Research Institute, University of Twente, P.O.Box 217, 7500AE Enschede, The Netherlands

ABSTRACT: A hardware implementation of a Backpropagation (BP) feedforward neural network has been designed. The tool was proposed for reflectometric measurements integrated together with photosensor arrays. The intelligent reflectometric sensor is being implemented in a multi-chip-module approach. A logarithmic input transformation is applied for easing the misalignment and parameter scatter correction. It also allows for easy ratio calculation by subtraction for normalisation with the reference value. The neural network was designed for complexities up to 100 inputs, 30 hidden neurons and 5 outputs. The digital building blocks (neurons) utilise a logic approximation of the sigmoid nonlinearity and the possibility of weight scaling. These hardware solutions result in a simultaneous area reduction and speed gain, at the cost of slightly decreased performance. Simulations of the proposed neural system prove applicability for evaluation of optical measurements. were performed for reflectometric and ellipsometric data thin porous layers. Hardware simulations showed good correspondence to the optimum-case neural software simulations.

Keywords: neural networks, spectroscopic measurements, photosensor arrays, digital multi-chip module

1. MOTIVATION

In the framework of the EU Copernicus 'RESPECT' project our goal was to design a demonstrative setup for real-time evaluation of spectroscopic measurements. We chose the Backpropagation type neural network as universal function mapping tool as our candidate for the task. Firstly computer simulations of such networks must have been carried out. Since a possible hardware realisation was concerned the reflectometric spectral analysis was chosen.

First of all reflection spectra of simple material structures were generated by optical simulation[1]. We used thin electrochemically etched porous silicon layers modelled by a mixture of silicon and air. Porosity is the percentage of the air in the layer. Effective porosity was changed by atmospheric oxidation[1].

The database created this way was used for neural network training. In neural theory for the determination of the number of necessary neurons there is still not reliable analytical method. Therefore simulations using different number of neurons were performed and finally 30 hidden neurons were taken into the only hidden layer. The threat of overtraining (i.e. spiky mapped function surface due to too numerous hidden neurons) was negligible as the training data bases contained 3000-5000 sample vectors. When our trained network was used in the recall mode with real-world samples which were not seen previously by the network, we got estimated material parameters in good correspondence with the parameters yielded from independent evaluation (Spectroscopic Ellipsometry[1]) as it is seen in Table I. The encouraging results reinforced our decision to design a hardware implementation for the same purpose.

Table I: Comparison of neural network approximations of material parameters from reflectometric spectra with independent evaluation method. The eventual weak correspondences are due to optical model imperfections and are being corrected. (aox: 300°C, box: 600°C oxidation in atmosphere)

Sample	956	956aox	956box	961	961aox	961box	970	970aox	970box
Calculated thickness (nm)	104	103	102	79	74	77	50	54	57
Neural approx. of thickness (nm)	107	115	91	82	84	69	49	43	58
Calculated porosity (%)	61	67	72	62	66	70	62	69	74
Neural approx. of porosity (%)	62	70	62	59	71	64	59	69	72
Sample	963	963aox	963box	969*	969*aox	969*box	969	969aox	969box
Calculated thickness (nm)	89	97	109	63	82	95	49	51	63
Neural approx. of thickness (nm)	86	83	56	69	66	47	49	49	49
Calculated porosity (%)	71	75	81	61	74	80	72	76	85
Neural approx. of porosity (%)	61	62	39	61	62	45	72	75	77

2. MODIFICATION OF THE BP ALGORITHM

During our simulations it was found, that after successful simulation some samples still have extreme high error values. This results from irregular behaviour of the non-linear function connecting the spectra and the material parameter values. Traditional BP algorithm uses all the samples with equal probability for training. The outcome is optimum performance in sense of root-of-mean-squares (RMSE). The statistically non-significant features of the input space, even though they can be characteristic, are missed from learning. Authors already reported attempts to use modified presentation probabilities for different samples in order to enhance learning speed[2,3]. However they suggested later abandoning the selected subset of training samples to avoid overtraining. A trade-off is to use a subset of the training database containing those samples, that have higher error than a selected threshold value. If that threshold is set continuously to a preset percentage of the highest occurring error level during training, the training subset can be selected also continuously. This avoids overtraining on one subset, and allows for shrinking the maximum error value. As it can be seen in Fig. 1 the statistically non-significant feature is much better learned by the modified algorithm. The W-C learning algorithm can be useful for example, when high confidence level of the expected maximum error level is needed, because the training parameter range is to be partitioned for successive neural parameter approximation. The introduced modification was used in simulations as it will be showed later.

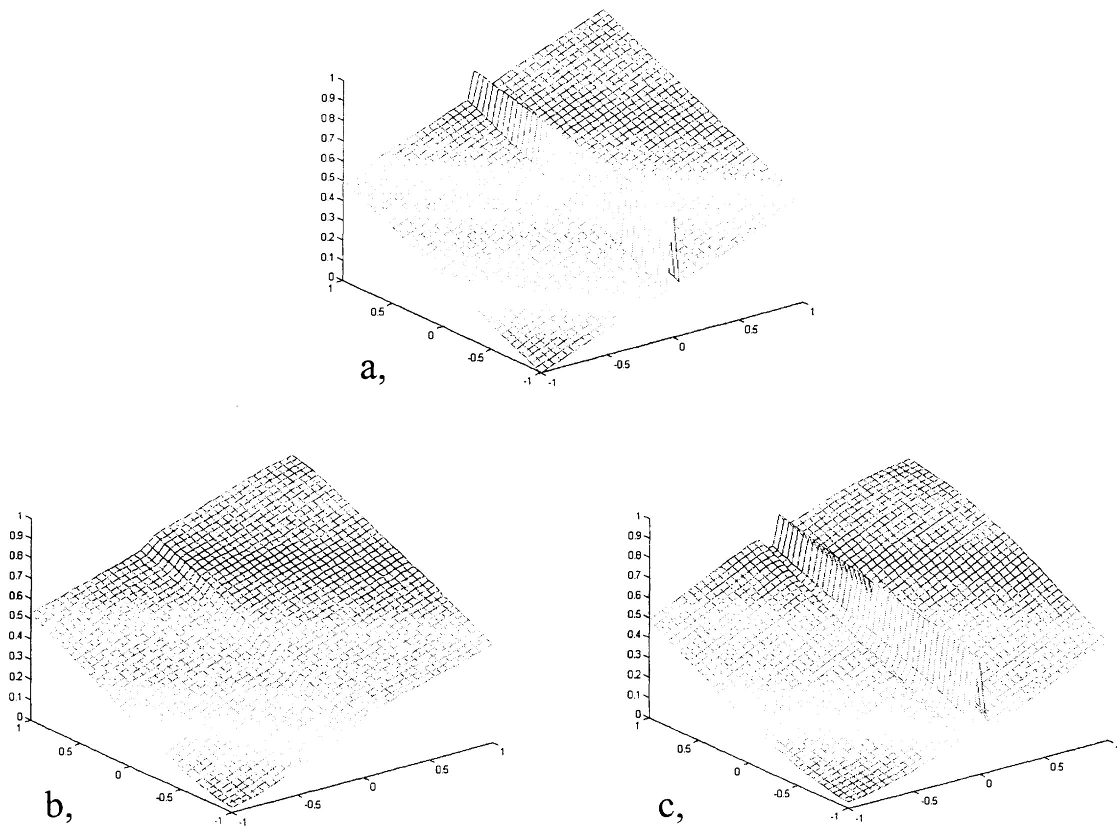

Fig. 1: Effect of worst-case (W-C) training on the approximation quality of a statistically non-characteristic feature. The target function is depicted in a); b) and c) are the attempts of neural networks for approximating it after training with traditional BP algorithm and the developed modification (W-C) respectively.

3. HARDWARE DESCRIPTION

After successful neural simulations a 100 input, 30 hidden neuron and 5 output neural reflectometric analysis system was designed. Input scaling and non-ideal behaviour concerns were taken into account as well as weight storage precision questions. After considering both analog and digital implementation[4-6] for easy redesign and precision reasons the digital approach was applied for implementation. This assumes A/D conversions from the analog optical input signals.

3.1 INPUT UNIT

The input unit of the reflectometry tool is simply an array of large photodiodes. Actually two identical arrays were designed for the system, one for measurement signals and the other one for reference signal. As the information in reflectometry is a relative intensity value, the measured light intensity signal (photocurrent) must be divided by the reference signal. In linear domain this implies complex computation in the digital circuitry.

For that reason a logarithmic input transformation is suggested. A chain of forward biased diodes are to transform the current signal into logarithmic voltage, which is readily lends itself for A/D conversion. That also eases multiplicative compensation for current deviations or fixed pattern noise. The corrective action is a pure addition. 10-bit A/D converters are used for digitalizing the inputs. The voltage margins are set so, that the difference between the two 10-bit values never exceeds the 8-bit range (i.e. 255). This way reasonable intensity resolution is obtained while 8-bit integer computational scheme is sustainable throughout the whole network.

3.2 NEURON ARCHITECTURE

The hidden neurons contain non-linear activation function. The output neurons are linear but otherwise their computational scheme is the same. As it can be seen in Fig. 2 the neurons have 8-bit signed integer weights. The first bit shows the sign and the rest seven bits store the value. The input of the neurons is an 8-bit unsigned integer. In each neuron there is a so-called shift value stored for virtually extending the weight resolution. Neural network simulations showed, that sometimes different neurons have the average of their weights in different order of magnitude. The shift value is used to scale the neurons weight sets separately. After a computer simulation the obtained weight values can be scaled to storable integer values this way.

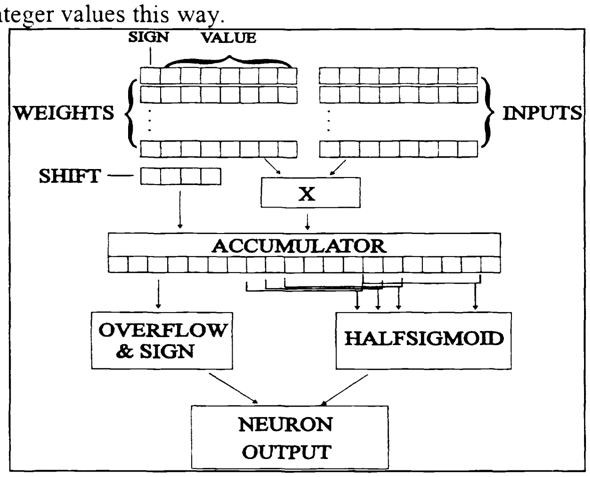

Fig. 2: Architecture of the digital neuron. The sign and overflow control signals are generated from the highest bit and those bits of the accumulator, which are above the shifted argument segment, respectively.

The non-linear Sigmoid function of the hidden neurons is implemented by combinational logic gate circuitry. The half-Sigmoid between 0 and 7.9375 independent values was mapped to the 127-255 8-bit integer range. The independent values were multiplied by 16 in order to fill the 0-127 integer range. The function values were rounded to integer values. The scaled integer half-sigmoid was approximated by random combinational logic circuits. The optimalisation of the random logic circuits were performed by using a genetic algorithm. The minimizable cost function for the genetic algorithm was the sum of differences for all the 128 input combinations. Finally the algorithm found a combinational logic, which produces a monotonous stepwise approximation of 16 steps of the half-Sigmoid. The gain is in area consumption and therefore speed enhancement due to parallelism. If a 256x8 bit RAM look-up table should have been added to each neuron, it could incredibly enlarge the neuron area. The solution in which a commonly used look-up table was used could result in saving silicon area, but also a reduction in speed. The half-Sigmoid contains about 200 transistors, and can be added to each hidden neuron. The silicon area saving can be significant in case of few weights, i.e. if the size of the weight memory is not the main part of the neuron size. It is inherently fast too, since it is asynchronous. The next section proves the applicability of the suggested simplifications.

4. COMPARISON OF SYSTEM SIMULATION TO COMPUTER SIMULATION

The neural system was fully simulated from input scaling and transformation to weight treatment and resolution matters. Weights were obtained from computer simulated training of a BP neural network. The training spectra were generated by a software using optical modelling. The logarithmic input transformation and appropriate scaling to the 8-bit integer range was done as a preparation of the training data. Computer simulation of an ideal (floating-point calculation and ideal sigmoid) neural network was performed using the given weights to approximate material parameters in a one layer porous silicon structure, as it was done in Section 1.

The same weights were scaled for each neuron into the 7-bit signed integer range, and the scaling factors were also determined. The system parameters were then used in the system simulation. The results are summarised in Table II. It is clear, that the absolute error introduced by the compromises in the designed hardware is acceptable. For some applications it can be regarded as good. Only the NNNbox, highly oxidized samples are badly approximated, probably due to the simple, two-parameter (thickness, effective porosity) optical model. Silicon, silicon-oxide, air model would be better.

It was also tested by computer simulation how positively the scaling affects the performance. Simulations were carried out without different scaling factors for individual neurons. When using more bits to represent the previously got floating-point weights, naturally we got approximation values closer to the floating-point case. According to our recall results on the training set the neuron-scope weight scaling affects the performance as if we had used 10-bit weight representation without scaling. The silicon area consumption is however different. In neuron-scope weight scaling we used eight more bits in each neuron to store the shift value, while 10-bit weight storage would have resulted in two hundred additional bits to store, and would have enlarged the fixed-point computation circuitry as well.

Table II: Results coming from the modified worst-case algorithm using global region and sub-region training data, and the corresponding simulated hardware responses (IC). The simulated IC responses are in good correspondence with the neural software simulations. (aox: 300°C, box: 600°C oxidation in atmosphere)

Sample	thickness [nm]				porosity [%]			
	global w-c	reg. w-c	reg. w-c(IC)	SR	global w-c	reg. w-c	reg. w-c(IC)	SR
956ref	101	107	109	105	68	62	62	61
956aox	92	115	116	103	71	70	69	67
956box	88	91	91	103	66	62	61	72
961ref	77	81	84	79	59	59	59	62
961aox	65	84	85	74	69	71	73	66
961box	59	69	71	77	61	64	64	70
970ref	41	49	50	48	61	59	61	60
970aox	56	43	43	54	68	69	70	69
970box	63	58	58	57	64	72	72	74
963ref	82	86	88	91	60	61	59	69
963aox	80	83	86	97	56	62	60	75
963box	45	56	52	109	8	39	43	81
969*ref	63	69	69	79	60	61	61	67
969*aox	59	66	65	82	66	62	63	74
969*box	43	47	44	95	39	45	44	80
969ref	56	49	50	49	64	72	72	72
969aox	53	49	49	51	68	75	76	76
969box	43	45	49	63	74	75	78	84

5. CONCLUSION, RECOMMENDATIONS

It was shown in our work, that compromises in digital design can make neurons more area-effective, while decrease the performance only slightly. Combinational logic function approximation prove to be a good alternative for look-up tables both in silicon area and speed aspects. Neuron-scope weight scaling is also an effective means of economic design. A digital neuron using the sketched ideas and having one hundred 8-bit weights was designed in the ES2 0.7μm standard CMOS silicon process. The achieved area is $1mm^2$, half of which is the size of the static RAM block. The chip is under fabrication.

The neuron chip testing will show the real practical value of our work. We furthermore recommend considering combinational logic utilisation for function approximations. Genetic algorithms can be effective in finding suitable approximators.

ACKNOWLEDGEMENTS

The authors thanks the EC for financing their efforts by the Copernicus 'RESPECT' project (No. CIPA-CT94-0131).

REFERENCES

1. M. Fried, T. Lohner, O. Polgár, P. Petrik, É Vázsonyi, I. Bársony, J. P. Piel and J. L. Stehle 1996, Characterization of different porous silicon structures by spectroscopic ellipsometry, Thin Solid Films v.276, pp. 223-227
2. Cho, S-B. and Kim, H. J. 1993. Rapid Backpropagation Learning Algorithms. Circuits, Systems and Signal Processing, Vol. 12, No. 2, pp. 155-175
3. Elhadi, N. and Cséfalvay, K. 1993. Backpropagation with Improved Learning Speed. Journal on Communications (A publication of the Scientific Society for Telecommunications, Hungary), Vol. XLIV., May, (1993).
4. Ienne, P., Cornu, T. and Kuhn, G. 1996. Special-Purpose Digital Hardware for Neural Networks: An Architectural Survey, Journal of VLSI Signal Processing Systems, vol. 13, pp. 5-25
5. Masa, P., 1994. NeuroClassifier - analog VLSI neural network for very high speed pattern classification, PhD. Thesis, University of Twente, reading
6. Wawrzynek, J., Asanovic, K. and Morgan, N. 1993. The Design of a Neuro-Microprocessor, IEEE Transactions on Neural Networks, VOL. 4, NO. 3, pp. 394-399