MCIT Modeling, control & information technologies

Modeling Nodes and Cells of Neuron-Equivalentors as Accelerators of Equivalental-Convolutional Self-Learning Neural Structures

https://doi.org/10.31713/MCIT.2024.077

Vladimir Krasilenko
Vinnytsia National Agrarian University
VNAU
Vinnytsia, Ukraine
krasvg@i.ua

Diana Nikitovich
Vinnytsia National Technical University
VNTU
Vinnytsia, Ukraine
diananikitovych@gmail.com

Alexander Lazarev
Vinnytsia National Technical University
VNTU
Vinnytsia, Ukraine
krasvg@i.ua

Abstract— **In the paper, we consider the urgent need to create highly efficient hardware accelerators for machine learning and deep convolutional neural networks (CNNs), for associative memory models, clustering, and pattern recognition. We show a brief overview of our related works the advantages of the equivalent models (EM) for describing and designing bio-inspired systems. The capacity of neural net on the basis of EM and of its modifications is in several times quantity of neurons. Such neuro-paradigms are very perspective for processing, clustering, recognition, storing large size, strongly correlated, highly noised images and creating of uncontrolled learning machine. And since the basic operational functional nodes of EM are such vector- matrix or matrix-tensor procedures with continuous- logical operations as: normalized vector operations "equivalence", "nonequivalence", and etc., we consider in this paper new conceptual approaches to the design of full-scale arrays of such neuron-equivalentors (NEs) with extended functionality, including different activation functions. Our approach is based on the use of analog and mixed (with special coding) methods for implementing the required operations, building NEs (with number of synapsis from 8 up to 128 and more) and their base cells, nodes based on photosensitive elements and CMOS current mirrors. Simulation results show that the efficiency of NEs relative to the energy intensity is estimated at a value of not less than 10^{12} an. op. / sec on W and can be increased. The results confirm the correctness of the concept and the possibility of creating NEs and MIMO structures on their basis.**

Keywords— accelerator; neural net; convolutional neural network; neuron-equivalentor; current mirror; vector-matrix procedure; equivalental model.

I. Introduction, Overviev, Analysis of Publications and Formulation of Work Goal

For many applications applied in the creation of biometric systems, machine vision systems are necessary to solve the problem of object recognition in images. The basis of most known methods and algorithms is to compare two different images of the same object or its fragment. Discriminant measure of the mutual alignment reference fragment with the current image, the coordinate offset is often a mutual two-dimensional correlation function. In paper [1, 2, 3] it was shown, that to improve accuracy and probability indicators with strong correlation obstacle-damaged image, it is desirable to use recognition methods based on mutual equivalently 2D spatial functions, nonlinear transformations and adaptive-correlation weighting. For the recognition and clustering of images, various models of neural networks are also used. Equivalental models (EM) of auto-associative memory (AAM) and hetero-associative memory (HAM) were proposed [2-6]. These EMs studies have shown, that these models allow the recognition of vectors with 1024 ÷ 4096 components and a significant percentage (up to 25-30%) of damage, at a network power that is 3 to 4 times higher than the number of neurons [3, 5, 6]. For of analysis and recognition should be solved the problem of clustering of objects. This previous clustering allows organizing proper automated grouping data, to cluster analysis, to evaluate on the basis of many signs each cluster, put a class label and improved subsequent learning procedures and classification. At the same time, knowing the significant advantages of EM whencreating on their basis improved neural networks (NNs),multiport AAM and HAM, there was a suggestion aboutthe possibility of modifying EM and MHAM for parallelcluster image analysis [6, 7, 8]. At the same time, an urgent task is to study a more general, spatially invariant(SI) equivalence models (SI EMs) that is more invariantto spatial displacements and the possibilities of its application for image clustering [7-9]. And the latter are basic operations in the most promising paradigms of convolutional neural networks (CNN) with deep learning [8, 9]. In our previous paper [8] questions of new possible ways of self-learning in such advanced models, explaining some important fundamentalconcepts of diverse associative recognition and understand the principles of the functioning of

biological NN-structures, perform modeling of processing processes [10], training and extraction of regularities in such models, and propose their implementation were considered. These questions were considered for bitmaps of multi-level images. In paper [9] we showed that the self-learning concept works with directly multi-level images without processing the bitmaps. In SI EM, we compute the spatially dependent normalized equivalence functions (SD_NEF) whose elements will correspond to the value of the normalized equivalence of the fragment of the input image X and one of the selected fragments from the training matrix. For implementation ESLCNS [9, 10], we need certain new or modified known devices capable of calculating normalized spatial equivalence functions (NSEqFs) with the necessary speed and performance. Such specialized devices by authors of papers were previously called "image equivalentor". There are known connections of equivalent functions with correlation functions that make it possible to calculate NSEqFs. Thus, the image equivalentor is itself a doubled correlator or a doubled convolver. In paper [8, 9] we showed models for the recognition and clustering of images that combine the process of recognition with the learning process. For all known convolutional neural networks, as for our equivalence models, it is necessary to calculate the convolution of the current fragment of the image in each layer with a large number of templates which are used, selected or formed during the learning process. But, as studies show, large images require a large number of filters to process images, and the size of the filters can also be large. Therefore, the problem of increasing the computing performance of hardware implementations of such CNNs is acute. It should be noted that the accuracy of calculations, especially for large filter sizes and a large dynamic range (8 bits) of halftone images, is required to make the correct decisions when determining neuron-winners. The last decade was marked by the activation of works aimed at the creation of specialized neural accelerators, which compute the function of comparing two 2D arrays and using the operations of multiplication and addition-accumulation. But as our experiments show, our models also allow the construction of ESLCNS. Therefore, in this paper, usingour approaches to designing one-dimensional neuron- equivalentors, we consider the structure of the neuron- equivalentor, generalized for processing 2D arrays.

II. Presentation of the main research results

A. Design of the main unit of ESLCNS

The Fig, 1 shows the block-diagram of the main unit of ESLCNS. The matrix **X** forms a certain number of convolutions in the form of matrices e using a set of defined filters-templates **W** which, in our case, are multilevel values, in contrast to the binary ones we used earlier. Thus, we compare each filter with a current fragment in the matrix **X**. As a measure of the similarity of the fragment of the matrix **X** and the filter the equivalent measures of proximity or other measures such as a histogram can be used. Thus, we compare for each filter similar fragments in the matrix. Fig. 1 shows the new structure of our proposed system, allowing parallel, with a high rate, equal to the speed of selection from the processed image of its shifted current fragment, to compute a set of stream components (elements) immediately one-cycle all the equivalents convolutions of the current fragment with the corresponding filters. It consists of a micro-display dynamically displaying current fragments, an optical node in the form of amicro-lens array (MLA) with optical lenses (not shown!) and a 2D array of neuron-equivalentors (**NEq**s)with optical inputs. Each **NEq** is implemented in a modular hierarchical manner and can consist of similar smaller sub-pixel, also 2D type, base nodes.

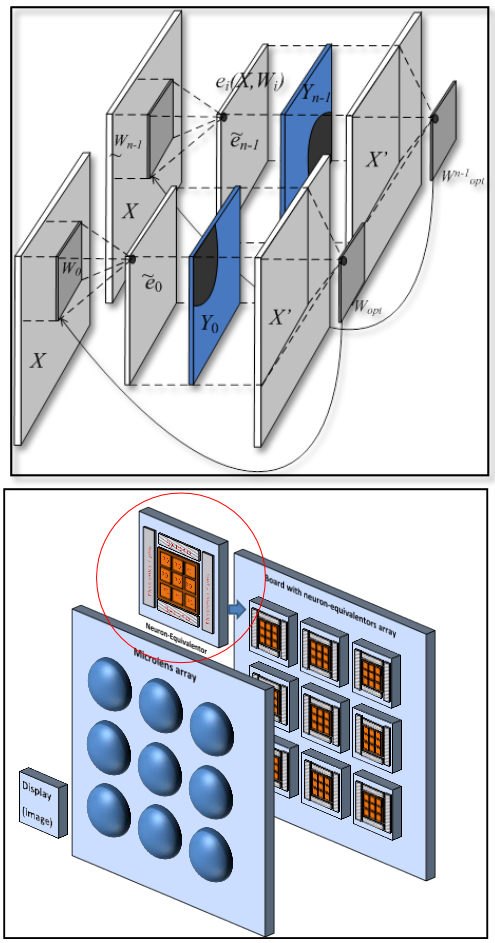

Figure 1. The structure of the basic unit of the ESLCNS (top). The neuron-equivalentors array (bottom).

The **NEq** has a matrix (ruler) of photo-detectors, on which a halftone image of the fragment is projected through the micro-lens array (MLA), and the number of electrical analog inputs equal to the number (number) of photo-detectors, to which by means of any known method: from the sample and hold device (SHD), from the analog memory, with subsequent conversion using a set of DACs, etc. the filter components are fed. These components are represented in the form of micro-currents. Each **NEq** has its own filter mask from a set of filters selected or formed by training. Thus, at the inputs of each **NEq** we have two arrays (vectors) of analog currents representing the compared current fragment and the corresponding filter-standard, and the output of the **NEq** is an analog current signal, nonlinearly transformed in accordance with the activation function and representing some measure of their similarity, proximity).

B. Design and Modeling of Nodes and Cells of Neuron-Equivalentors as Accelerators

In our case, this measure is a normalized equivalence (eq) and nonequivalence (neq), we can calculate them by averaging the component maxima and minima currents. Therefore, the base node, see Fig. 2-4, contains N two input counters of maximum and minimum currents and one normalizer on current mirrors, which forms two output signals corresponding to normalized eq and neq from two N-component vectors.

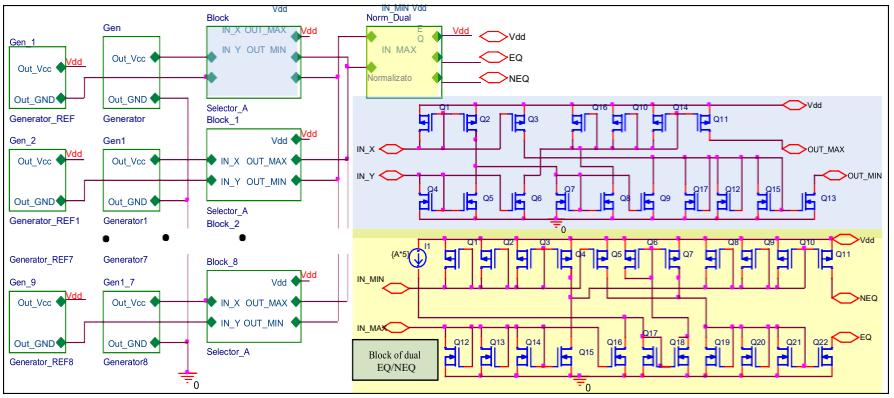

Figure 2. The basic unit for calculating the normalized Eq (NEq) by averaging the component peak and minima of currents on the basis of current mirrors and the schemes of the limited difference.

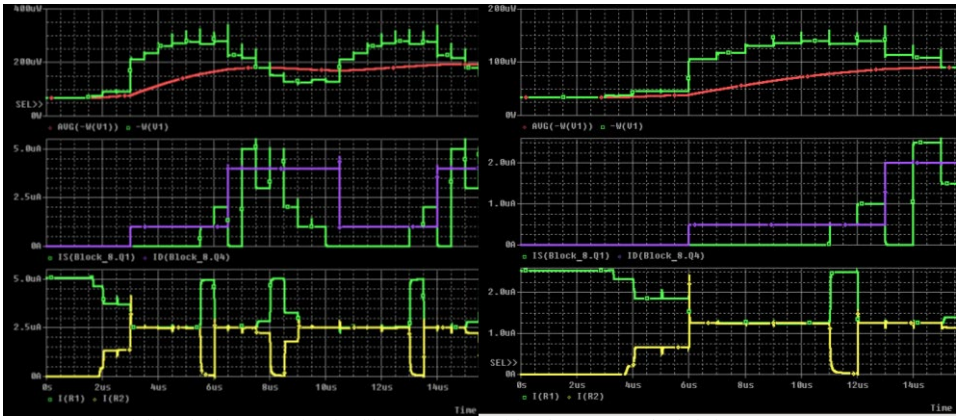

Figure 3. The results of modeling the base unit for the filter size 3x3 (with 9 inputs): on the left for current Imax=5μA, T=0.5μs, V=1.8V, P=200μW. On the right for current Imax=2.5μA, T=1μs, P=100μW. Red line shows power consumption, input (green) and reference (lilac) signals are showed on the middle graphs, on the bottom graph normalized eq (green) and neq (yellow) are showed.

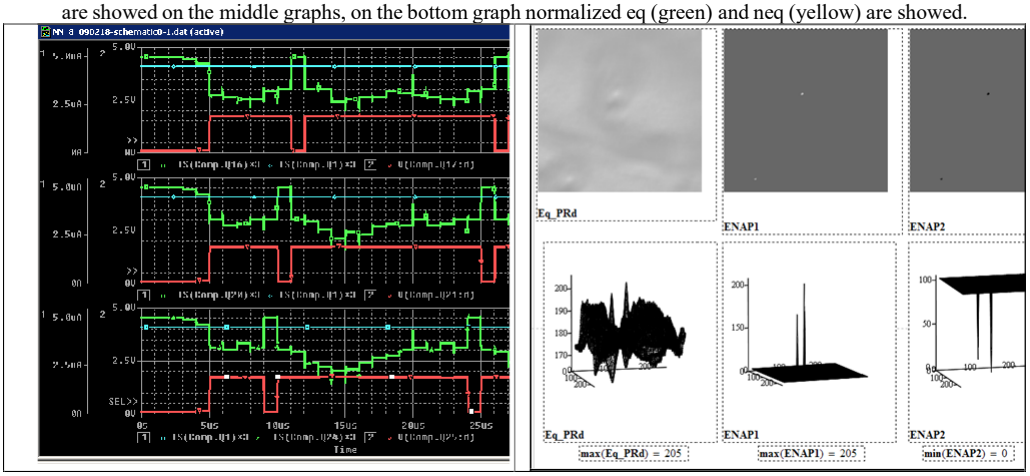

Figure 4. On the left: The result of a network simulation of 8 9-input NEs, a fragment of the successive activity of three neighboring NEs, green - current outputs, blue-additional threshold, red-NE outputs (voltage, potential). On the right: The Mathcad windows on which the module of the program with formulas and results of recognition of fragments on the image are shown, where in 2D and 3D from left to right: the computed NE equivalent, non-linear (after activation) equivalent, linear non- equivalent (part) functions.

The basic unit for calculating the normalized Eq (NEq) by averaging the component peak and minima of currents on the basis of current mirrors and the schemes of the limited difference is shown in Fig. 2. Sources of analog currents are shown as current generators for modeling in OrCAD. The dimension of the input vectors is 9, which corresponds to the filter size 3x3. The results of modeling this nonlinear transformation basic node are shown in Fig. 3 As can be seen from the diagram, the complete coincidence of the input vectors fed to the input of this node and changing over time is evidenced by the changes in the complementary output signals of equivalence and non-equivalence (increasing current to a maximum and a corresponding complementary decrease to a minimum!) at the corresponding moments of time (11 - 12 μs and others).The results of modeling the base unit for the filter size 3x3 (with 9 inputs) showed, that processing time is from 1μs to 0.1μs for currents Imax=5μA, consumption power is from 200μW to 50μW. In addition to simulating the base node on 9 inputs, we additionally synthesized a neuron-equivalent circuit having 8 such nodes, each of which compares 8 input vectors, resulting in a neural element circuit having 2 vector inputs of 64 dimensions. For a non- linear transformation, we used a node of a circuit, whichrealizes a piecewise linear approximation of the power- law activation function (auto-equivalence). The simulation results of 64 input NE with nonlinear output conversion showed that such a NE comparing the two 64 component vectors from the current signals provides good time characteristics and has a total power consumption of approximately 2mW, a low supply voltage, contains less than 1000 CMOS transistors with which summation circuits are implemented, limited subtraction and multiplication of analog currents on current mirrors. The simple build-up of nodes and additional introduction of normalizers for equalization and matching of signal levels in micro-nodes allow to increase the number and dimension of inputs and thus increase the dimension of filters. We designed and modeled a neuron-equivalentor for two 81-component input vectors, i.e., to implement convolution with a filter of dimension 9 × 9, by combining nine micro-nodes, namely neuron-equivalentors with two 9-component vector inputs. It has 2 analog input buses. And the results of its modeling are shown in Fig. 4. To verify the functioning of the developed neuron-equivalentors in the network, a neural mini-network of eight 9-input micro-nodes neuron-equivalentors was created, the simulation results of which also confirmed its correct functioning.

III. Discussion

Within the framework of this work, the set goal was achieved, namely, the principles of implementation of neuron-equivalents as accelerators and their basic cells were developed. The use of analog current reflectors as basic elements for the construction of all nodes, including the node for generating activation functions, neuron-equivalents made it possible to show and experimentally confirm the advantages and prospects of the proposed technical solutions. At the same time, some aspects important for the implementation of the proposed models and their schematic solutions and the further expansion of the spheres of their use remained unexplored. This includes the implementation of physical models and their testing, measurement of characteristics and indicators for the purpose of their comparative analysis with other concepts.

IV. Conclusions

Neuron-equivalentors have a processing-conversion time of 0.1-1μs, low supply voltages of 1.8-3.3V, minor relative computational errors (1-5%), small consumptions of no more than 1-2mW, can operate in low-power modes less than 100μW) and high-speed (10-20 MHz) modes. The efficiency of neuron-equivalentors relative to the energy intensity is estimated at a value of not less than 10 ^ 12 an. op. / sec on W and can be increased by an order of magnitude. The obtained results confirm the correctness of the chosen concept and the possibility of creating hierarchical macro neuron-equivalentors and MIMO structures based on them. They can become the basis for the implementation of CNN and self-learning biologically inspired devices with the number of such neuron-equivalentors equal to 1000, to realize the parallel calculation of equivalental convolutions with filter sizes up to 32* 32.

References

[1] Krasilenko, V. G., Saletsky, F. M., Yatskovsky, V. I., Konate, K., "Continuous logic equivalence models of Hamming neural network architectures with adaptive-correlated weighting," Proceedings of SPIE Vol. 3402, pp. 398-408 (1998).

[2] Krasilenko, V. G., Magas, A. T., "Multiport optical associative memory based on matrix-matrix equivalentors," Proceedings of SPIE Vol. 3055, pp. 137 – 146.

[3] Krasilenko V.G., Nikitovich D.V., "Experimental studies of spatially invariant equivalence models of associative and hetero-associative memory 2D images," Systemy obrobky informaciji, 4 (120), 113 –120 (2014).

[4] Krasilenko V.G., Nikolskyy, A. I., "The associative 2D-memories based on matrix-tensor equivalental models," Radioelektronika Inform. Communication, 2 (8), 45 –54 (2002).

[5] Krasilenko, V. G., Lazarev, A., Grabovlyak, S., "Design and simulation of a multiport neural network heteroassociative memory for optical pattern recognitions," Proceedings of SPIE Vol. 8398, 83980N-1 (2012).

[6] Krasilenko V. G., Lazarev, A., Grabovlyak, S., Nikitovich D.V., "Using a multi-port architecture of neural-net associative memory based on the equivalency paradigm for parallel cluster image analysis and self-learning," Proceedings of SPIE Vol. 8662, 86620S (2013).

[7] Krasilenko V.G., Nikitovich D.V., "Simulation of self-learning clustering methods for selecting and grouping similar patches, using two-dimensional nonlinear space-invariant models and functions of normalized "equivalence," Electronics and information technologies: collected scientific papers, Lviv: Ivan Franko National University of Lviv, Issue 6, pp. 98-110 (2015).

[8] Krasilenko V.G., Lazarev A.A., Nikitovich D.V., "Modeling and possible implementation of self-learning equivalence-convolutional neural structures for auto-encoding-decoding and clusterization of images", Proceedings of SPIE Vol. 10453, 104532N (2017).

[9] Krasilenko V.G., Lazarev A.A., Nikitovich D.V., "Modeling of biologically motivated self-learning equivalent-convolutional recurrent-multilayer neural structures (BLM_SL_EC_RMNS) for image fragments clustering and recognition", Proceedings of SPIE Vol. 10609, MIPPR 2017: Pattern Recognition and Computer Vision, 106091D (8 March 2018); doi: 10.1117/12.2285797.

[10] Krasilenko, V. G., Nikolskyy, A. I., Lazarev A.A., "Designing and simulation smart multifunctional continuous logic device as a basic cell of advanced high-performance sensor systems with MIMO-structure," in Proceedings of SPIE, 94500N (2015).