# Analog Neuromorphic Computing Enabled by Multi-Gate Programmable Resistive Devices

Vehbi Calayir, Mohamed Darwish, Jeffrey Weldon, and Larry Pileggi
Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

***Abstract*—Analog neural networks represent a massively parallel computing paradigm by mimicking the human brain. Two important functions that are not efficiently built by CMOS technology for their practical hardware implementations are *weighting* for synapse circuits and *summing* for neuron circuits. In this paper we propose the use of tunable analog resistances, such as multi-gate graphene devices, to efficiently enable these two functions. We design and demonstrate a complete analog neuromorphic circuitry enabled by such devices. Simulation results based on Verilog-A compact models for graphene devices confirm its functionality. We also provide experimental demonstration of our proposed graphene device along with projected circuit performance based on scaling targets. Our proposed design is suitable not only for the device example shown in this paper, but also for any beyond-CMOS technology that exhibits similar device characteristics.**

## I. Introduction

The human brain is a powerful computing system that performs information processing quite efficiently via its massively parallel neural mechanism. Neuromorphic circuits attempt to mimic this neural mechanism in the human brain via locally-coupled artificial neurons and synapses in order to overcome some bottlenecks of the traditional von Neumann machines, especially for computationally-intensive applications such as image processing and pattern recognition [1]-[3].

A conceptual neuromorphic computing architecture that can be used for both feed-forward and recurrent (feedback) networks is depicted in Fig. 1. In this model the outputs of the neighboring neurons are weighted with corresponding synaptic weights. Then these weighted signals are summed together in the neuron circuit to generate its own state. Finally, the neuron circuit takes this summation and generates its output based on an activation function that generally corresponds to a variant of the sigmoid function.

Two key challenges that require significant power and area resources when implemented with CMOS technology are to efficiently represent synaptic weights and sum neural signals coming from neighboring synapses. For example, the proposed CMOS-based designs in [4]-[5] attempt to use one digital-to-analog converter (DAC) and one variable gain multiplier for each artificial synapse, and one differential summing amplifier to represent each artificial neuron.

Although CMOS is inefficient for implementing analog neuromorphic circuits, several emerging technologies have been recently proposed for such purposes by exploiting their unique features [2]-[3]. In such implementations weighted signals are in current domain and various coupling mechanisms (e.g., thermal [2] and magnetic [3]) have been utilized to efficiently sum such current-mode signals and then generate a voltage-mode output signal. In this paper we propose to construct a fully-functional analog neural network using multi-gate programmable graphene devices. Different from the proposed designs in [2]-[3] this architecture represents couplings between neighboring neurons in voltage domain, and uses only electrical signals for both weighting and summing functions, thereby not requiring any special non-electrical properties for the enabling device.

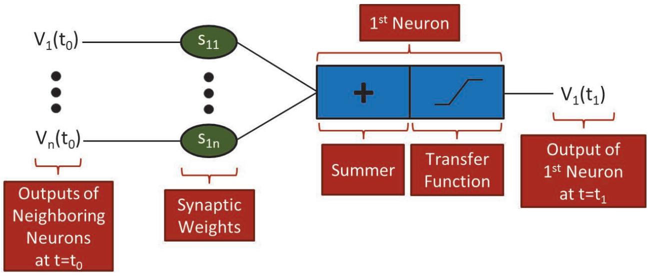

Fig. 1. A generalized building block for analog neuromorphic circuitry.

## II. Multi-Gate Tunable Resistive Device Example

Graphene is a two-dimensional material consisting of a single-atom thick layer of carbon that is arranged in a hexagonal lattice (see Fig. 2). With its 2D planar structure it has been shown to be compatible with traditional CMOS process [6]. Moreover, the charge carriers in graphene can move over great distances at a constant speed without scattering. This is similar to the behavior of photons travelling at the speed of light. In graphene the speed of charge carriers is slower than light by only a factor of 300 [7], thereby resulting in high saturation velocity. Graphene also has high carrier mobility and zero energy band-gap. Thus, it offers unique opportunities for future nanoelectronics such as high frequency applications [8]. In addition, the linear dispersion relation in graphene gives rise to relativistic behavior of charge carriers resulting in photon-like behavior and Klein tunneling [9].

Here we propose a novel graphene device configuration that could be used to represent artificial neurons and synapses. The number of charge carriers in a graphene ribbon can be controlled through local gating [10] that allows the gradual resistance change of the device. This is because the electric field between the gate and graphene ribbon due to applied gate voltage attracts electrons in graphene, thereby altering its conductivity. By means of this feature and zero energy band-gap offered by this particular material, a resistor string can be built using multiple gates over a single continuous graphene ribbon, as illustrated in Fig. 3. Although the gates effectively dope the graphene creating multiple junctions, conductivity is maintained due to Klein tunneling at the junctions [9]. The

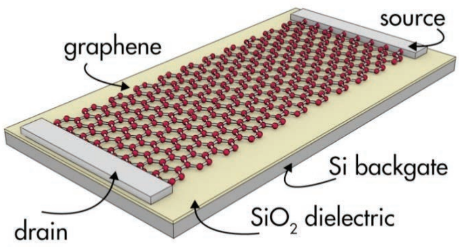

Fig. 2. A graphene device [11].

This work was supported by the Systems on Nanoscale Information fabriCs (SONIC), one of six centers supported by the STARnet phase of the Focus Center Research Program (FCRP), a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

978-3-9815370-4-8/DATE15/©2015 EDAA

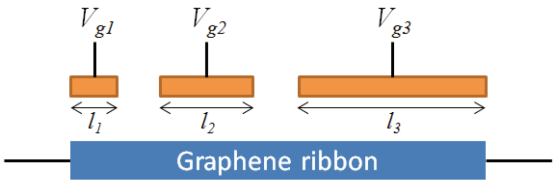

Fig. 3. Cross-section drawing of the multi-gate graphene resistance. Each gate controls the conductivity of the corresponding area underneath it. $V_{g1}$, $V_{g2}$ and $V_{g3}$ represent voltage signals applied to these gates. The number of gates over graphene ribbon can be easily increased.

contribution of each gate on the resistance level can be adjusted for specific purposes by simply altering the corresponding gate length (i.e., $l_1$, $l_2$ and $l_3$ in the figure). Such a multi-gate graphene device performs the DAC and summing functions that can be utilized for implementing artificial neurons and synapses without requiring any other data converters or CMOS amplifiers.

## III. Proposed Neuromorphic Computing Circuits

The neuron-synapse building block for our proposed architecture is shown in Fig. 4. Each artificial neuron is constructed as one pull-up ($R_{PU}$) and one pull-down ($R_{PD}$) graphene device with multiple gates. In combination they function similar to an *analog buffer* that produces intermediate values between high and low voltage levels. Each gate in these two devices has an equal impact on the total resistance (i.e., each having the same length, referring to Fig. 3), and the number of gates per device in the neuron circuits is equal to the number of neighboring neurons. As such, these devices enable an efficient summing mechanism for weighted neural signals coming from neighboring synapses via equally-sized gates.

Each synapse between neighboring neurons is categorized into two components: i) *excitatory synapse* and ii) *inhibitory synapse*, as shown in Fig. 4. Each of these components is represented by two multi-gate programmable resistances. Each gate in both devices controls a binary-weighted resistor based on the corresponding area on graphene ribbon (e.g., $l_1=x$, $l_2=2x$, and $l_3=4x$, referring to Fig. 3). The number of control bits, $i$, can be adjusted based on the requirements for the network function where the number of allowable resistance values is equal to $2^{i+1}-1$, because each synapse consists of two components. Generally 15-31 different synaptic values are enough for a wide range of neural network applications [12]-[13]. This device configuration enables a compact D/A (digital-to-analog) conversion for digitally-controlled synaptic weights.

The total resistance of two devices in each synapse component is always constant since binary bits applied to them are complementary to each other (see Fig. 4). Hence the couplings between neuron output and synaptic weights are linear as given by:

$$V_{ex\_w} = V_{neu\_out} \frac{R_{ex\_w}}{R_{total}} \quad \textit{for excitatory synapse} \qquad (1)$$

$$V_{in\_w} = V_{neu\_out} \frac{R_{in\_w}}{R_{total}} \quad \textit{for inhibitory synapse} \qquad (2)$$

where $V_{neu\_out}$ is the output of the corresponding neuron; $R_{ex\_w}$ and $R_{in\_w}$ denote excitatory and inhibitory synaptic weights, respectively; $V_{ex\_w}$ and $V_{in\_w}$ are weighted voltage signals corresponding to excitatory and inhibitory components, respectively; and

$$R_{total} = R_{ex\_w} + R_{ex\_w}' = R_{in\_w} + R_{in\_w}' \qquad (3)$$

where $R_{ex\_w}'$ and $R_{in\_w}'$ are complementary versions of $R_{ex\_w}$ and $R_{in\_w}$, respectively.

The excitation component of each synapse circuit is activated by increasing $R_{ex\_w}$ via digital control bits (i.e., $b_{ex,1}$-$b_{ex,i}$ in Fig. 4) when neighboring neurons are coupled to each other via a *positive* synaptic weight (i.e., two neighboring neurons, each representing an image pixel, that are more likely to be of the same color, either white or black). The inhibition component is activated in the same way as the excitation component (i.e., via $b_{in,1}$-$b_{in,i}$ in Fig. 4) when the relationship between neighboring neurons is represented by a *negative* synaptic weight (i.e., two neighboring neurons, each representing an image pixel, that are more likely to be of the opposite colors). A synaptic weight can be either positive or negative, so the corresponding synapse component is deactivated by setting $R_{ex\_w}$ or $R_{in\_w}$ to its lowest possible value (i.e., either $b_{ex,1}$-$b_{ex,i}$ or $b_{in,1}$-$b_{in,i}$ are set to 0).

An interesting property of this proposed architecture is that couplings between neighboring neurons are voltage-mode signals instead of current-mode signals, in contrast to other proposed resistive networks [2]-[3]. This is dictated by the gates being controlled by voltage-mode signals.

It is important to note that inhibitory synapses are connected to the gates of pull-up device in the neuron circuits, while excitatory synapses are connected to the gates of pull-down device, as shown in Fig. 4. This is because weighted signals coming from artificial synapses increase the resistance of the corresponding device. Therefore, when excitation becomes more dominant as compared to inhibition, the neuron circuit generates an output level that is closer to the supply voltage, $V_{supply}$ in Fig. 4. In contrast, when inhibition is more dominant, the neuron circuit produces an output that is closer to 0V (ground).

## IV. Circuit Simulations

To evaluate the potential of our proposed neurocomputing architecture in Fig. 4 we have created a compact circuit simulation model in Verilog-A for the graphene device configurations based on device measurement data [10]. The relationship between graphene resistance and applied gate

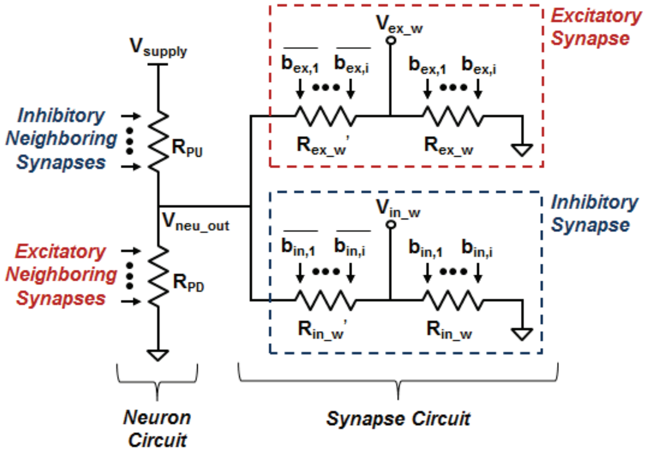

Fig. 4. Proposed neuron and synapse circuits based on multi-gate programmable resistive devices. $b_{ex,1}$-$b_{ex,i}$ and $b_{in,1}$-$b_{in,i}$ represent $i$-bits binary numbers for excitatory and inhibitory synaptic weights, respectively.

voltage is modeled as a saturation function with two cut-off voltages (i.e., $V_{cut\_min}$ and $V_{cut\_max}$). This function outputs the lowest resistance value ($R_{min}$) when applied gate voltage is less than $V_{cut\_min}$, and the highest resistance value ($R_{max}$) when applied gate voltage is higher than $V_{cut\_max}$. When it is in-between these two voltage levels the device resistance increases linearly with applied gate voltage (from $R_{min}$ to $R_{max}$).

In our circuit simulations we used $V_{supply}=1V$, $V_{cut\_min}=0.1V$, $V_{cut\_max}=0.9V$, and $R_{max}/R_{min}=100$ which is achievable with today's technology [10]. This R-V characteristic can be attained by means of a properly-selected reference voltage corresponding to the ground in our circuits (e.g., around -3V for the device example in [10]), or a carefully-tuned back-gate/substrate voltage, or both. We connected 1pF capacitor at the output of each neuron circuit to model device parasitics.

The devices used in the neuron circuits have nine equally-weighted gates (i.e., $l_1=l_2=\ldots=l_9$) so that each neuron has nine local connections in the network. The devices used in the synapse circuits, however, have three binary-weighted gates (i.e., $l_1=2l_2=4l_3$) that allow 15 different synaptic weights that can be defined in the model (considering both excitatory and inhibitory synapses). The device resistances in the synapse circuits are designed to be 100 times greater than those in the neuron circuits in order to prevent significant deviations in the neuron output voltages from their actual values due to fanout.

As a first example we constructed a 5-neuron associative memory based on the architecture proposed in Fig. 4. Using pattern examples [1 0 1 1 0] and [1 0 1 0 1] as memorized patterns by programming excitatory and inhibitory synapses accordingly using *Hebbian Rule* [14], we evaluated the convergence of this small network when other patterns were provided as inputs. Fig. 5 shows a pattern recognition example. Even though there is 40% distortion in the initial input pattern [0 1 1 1 0], associative memory successfully recalls stored pattern that most closely resembles to applied input pattern.

We also built 20-pixel gray-scale example for larger neuromorphic circuits with local interconnections among artificial neurons (e.g., 9 connections for each neuron) based on the architecture in Fig. 4. For illustration we show here the same device simulation models to evaluate an associative memory with 20 neurons. The bit patterns shown in Fig. 6 are examples of memorized patterns that are programmed into the neural network circuit model via synaptic weights.

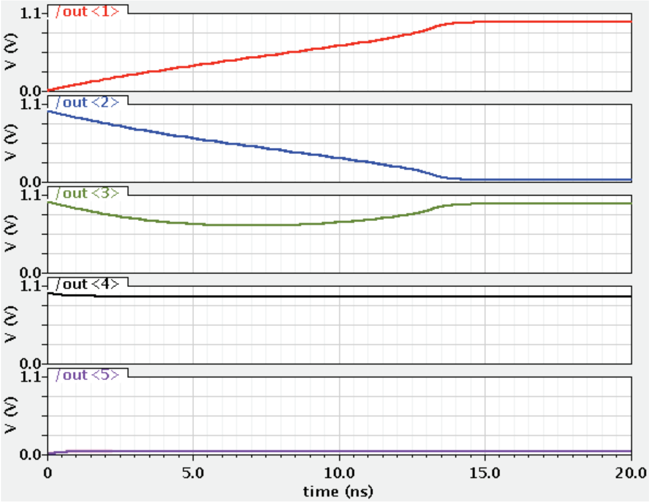

Fig. 5. Pattern recognition example based on a 5-neuron system. Associative memory fully recognizes the pattern [1 0 1 1 0] despite 40% distortion in the initial input pattern.

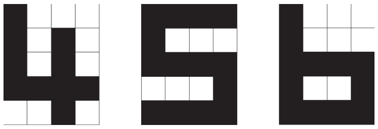

Fig. 6. Memorized bit patterns for a 20-neuron network.

As an initial input pattern we used a 28% distorted version of the bit pattern '5' with several gray-scale pixels, and our associative memory successfully recovers the bit pattern '5' as illustrated in Fig. 7. Importantly, the results here indicate that our proposed neuromorphic circuits scale well with the number of bits as long as the nearest neighbor coupling is sufficiently accurate. Moreover, all the results presented here match the ones in [2], thus confirming correct functionality.

Next we highlight the use of the same 20-neuron network for other image processing applications such as edge and line detections. In order to program excitatory and inhibitory synapses we used the cloning templates for edge and line detection applications given in [13], [15]. Fig. 8 presents our circuit simulation results for these applications.

## V. Prototype Device & Projected Circuit Performance

A prototype device was fabricated using a CMOS compatible process. First photolithography was used to define the contacts which were formed via physical vapor deposition of Cr/Au: 5nm/25nm. Next graphene was patterned using photolithography and etched using oxygen plasma etching to define channels with required length and width. After annealing the graphene to remove any residual photoresist, the gate oxide stack was deposited using a 2nm aluminum seed followed by 8nm aluminum oxide deposition using atomic layer deposition. The gate metals were then defined using electron beam lithography and deposited using evaporation. The false image color of the active area for the fabricated device is shown in Fig. 9.

This device consists of three top gates, the lengths of which increase in a binary fashion, along with a back gate. Each top gated section acts as a variable resistance that can be controlled by a combination of applied top and back gate voltages. The back gate voltage is used to bias the device in the resistance range of interest, and the top gates are used to control the number of charge carriers under the gates, thereby effectively modulating the resistance value within the range of interest.

Measurement results are provided in Fig. 10. Although this prototype used for characterization is a larger device (1μmx10μm) than we would expect to use in our proposed architecture, it does demonstrate the expected behavior. This first-of-its-kind device provides the monotonic behavior as a function of applied gate voltage, and shows the unique capability of graphene to enable the desired multi-gate programmable resistance in a compact and efficient way. With further development and scaling of this technology, highly-efficient neuromorphic circuit components could be fabricated and integrated together to construct a fully-functional system.

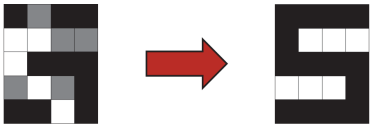

Fig. 7. The initial input pattern for a 20-neuron based associative memory (left) and the output pattern produced by this associative memory (right).

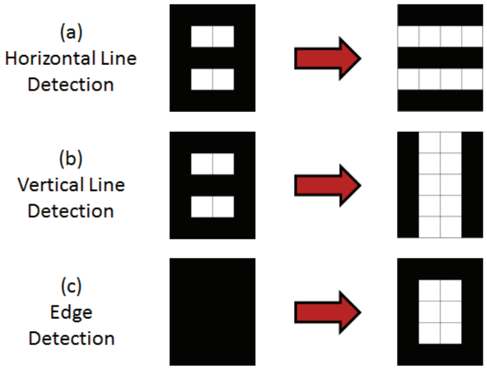

Fig. 8. Circuit simulation results for our proposed neurocomputing circuits. Input patterns (left) and output patterns (right).

It should be noted that this controlled resistance based on gating of a zero band-gap graphene ribbon would not be possible with a MOSFET device due to the inherent threshold voltage of the latter.

With proper selection of back gate voltage, the minimum device dimensions possible with today's lithography could be used for the proposed graphene devices while still attaining the required high-to-low resistance ratio (e.g., 100 used in our circuit simulations). A sheet resistance of approximately 1kΩ can be achieved for a 0.1nm thick graphene ribbon using today's technology [16]. A minimum size device can be used in the neuron circuits, and larger devices in the synapse circuits (e.g., 100x larger). Our simulations indicate that with the device width of 10nm and varying lengths of 10nm to 1μm for the minimum-size devices used in the circuitry the proposed implementation provides *more than three orders of magnitude improvement in both power and area* when compared to a CMOS-based design fabricated in 65nm.

## VI. Conclusion

In this paper we propose and design a complete analog neuromorphic circuitry based on the multi-gate programmable resistances that provide a natural summing and compact D/A conversion for the implementation of artificial neurons and synapses. The operation of the proposed architecture has been confirmed using simple gray-scale pattern recognition and image processing examples. The fabrication process and measurement results for enabling device have been provided to anticipate scaling targets required for a highly-efficient network realization. This research is intended to provide insights into the development of new devices with intriguing capabilities that could enable feasible implementations of analog neurocomputing circuits.

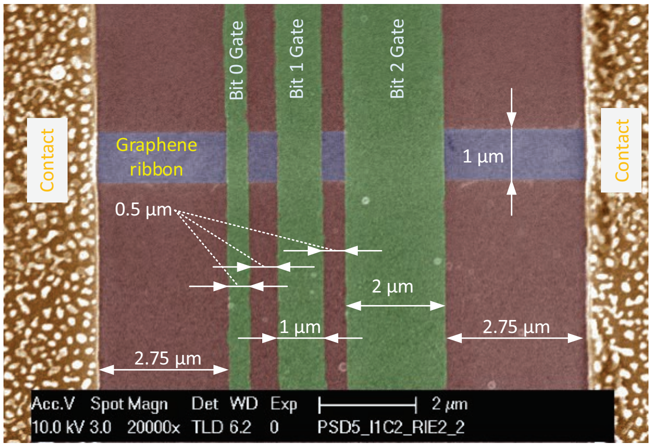

Fig. 9. The false color image of the active area for three-gate prototype device showing device dimensions.

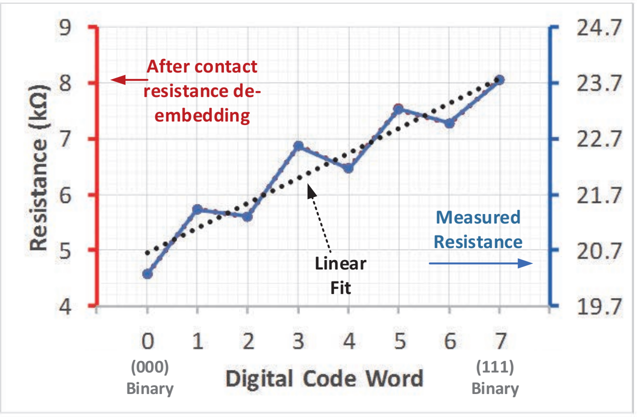

Fig. 10. Resistance measurements for the fabricated device. Resistance change is 3.5 kΩ with a 15.7 kΩ contact resistance.

## References

[1] V. Calayir and L. Pileggi, "Fully-digital oscillatory associative memories enabled by non-volatile logic," *International Joint Conference on Neural Networks*, pp. 1-6, Aug. 2013.

[2] V. Calayir, T. Jackson, A. Tazzoli, G. Piazza, and L. Pileggi, "Neurocomputing and associative memories based on ovenized aluminum nitride resonators," *International Joint Conference on Neural Networks*, pp. 1-8, Aug. 2013.

[3] V. Calayir and L. Pileggi, "All-magnetic analog associative memory," *IEEE International New Circuits and Systems Conference*, pp. 1-4, June 2013.

[4] P. Kinget and M. S. J. Steyaert, "A programmable analog cellular neural network CMOS chip for high speed image processing," *IEEE Journal of Solid State Circuits*, vol. 30, no. 3, pp. 235-243, March 1995.

[5] J. M. Cruz and L. O. Chua, "A 16x16 cellular neural network universal chip: the first complete single-chip dynamic computer array with distributed memory and with gray-scale input-output," *Analog Integrated Circuits and Signal Processing*, vol. 15, no. 3, pp. 227-237, March 1998.

[6] M. C. Lemme, T. J. Echtermeyer, M. Baus, and H. Kurz, "A graphene field-effect device," *IEEE Electron Device Letters*, vol. 28, no. 4, pp. 282-284, March 2007.

[7] R. M. Westervelt, "Graphene nanoelectronics," *Science*, vol. 320, no. 5874, pp. 324-325, Apr. 2008.

[8] Y.-M. Lin, K. A. Jenkins, A. Valdes-Garcia, J. P. Small, D. B. Farmer, and P. Avouris, "Operation of graphene transistors at gigahertz frequencies," *Nano Letters*, vol. 9, no. 1, pp. 422-426, Jan. 2009.

[9] P. E. Allain and J. N. Fuchs, "Klein tunneling in graphene: optics with massless electrons," *The European Physical Journal B*, vol. 83, no. 3, pp. 301-317, Oct. 2011.

[10] B. Ozyilmaz, P. Jarillo-Herrero, D. Efetov, and P. Kim, "Electronic transport in locally gated graphene nanoconstrictions," *Applied Physics Letter*, vol. 91, no. 19, pp. 192107, Nov. 2007.

[11] J. Hedberg [Online]. Available: www.jameshedberg.com/img/samples/.

[12] T. Pfeil, T. C. Potjans, S. Schrader, W. Potjans, J. Schemmel, M. Diesmann, and K. Meier, "Is a 4-bit synaptic weight resolution enough? – constraints on enabling spike-timing dependent plasticity in neuromorphic hardware," *Frontiers in Neuroscience*, vol. 6, no. 90, pp. 1-19, July 2012.

[13] E. Raschman, R. Zalusky, and D. Durackova, "New digital architecture of CNN for pattern recognition," *Journal of Electrical Engineering*, vol. 61, no. 4, pp. 222-228, July-Aug. 2010.

[14] A. D. B. Delbem, L. G. Correa, and L. Zhao, "Design of associative memories using cellular neural networks," *Neurocomputing*, vol. 72, no. 10-12, pp. 2180-2188, June 2009.

[15] L. O. Chua and L. Yang, "Cellular neural network: applications," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 10, pp. 1273-1290, Oct. 1988.

[16] R. Murali, K. Brenner, Y. Yang, T. Beck, and J. D. Meindl, "Resistivity of graphene nanoribbon interconnects," *IEEE Electron Device Letters*, vol. 30, no. 6, pp. 611-613, June 2009.