Combining a Volatile and Nonvolatile Memristor in Artificial Synapse to Improve Learning in Spiking Neural Networks
Mahyar Shahsavari, Pierre Falez, Pierre Boulet

To cite this version:

HAL Id: hal-01368954
https://hal.archives-ouvertes.fr/hal-01368954
Submitted on 20 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Combining a Volatile and Nonvolatile Memristor in Artificial Synapse to Improve Learning in Spiking Neural Networks

Mahyar Shahsavari  
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRISAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France  
Mahyar.Shahsavari@ed.univ-lille1.fr

Pierre Falez  
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRISAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France  
Pierre.Falez@ed.univ-lille1.fr

Pierre Boulet  
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRISAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France  
Pierre.Boulet@univ-lille1.fr

ABSTRACT

With the end of Moore's law in sight, we need new computing architectures to satisfy the increasing demands of big data processing. Neuromorphic architectures are good candidates to low energy computing for recognition and classification tasks. We propose an event-based spiking neural network architecture based on artificial synapses. We introduce a novel synapse box that is able to forget and remember by inspiration from biological synapses. Two different volatile and nonvolatile memristor devices are combined in the synapse box. To evaluate the effectiveness of our proposal, we use system-level simulation in our Neural Network Scalable Spiking Simulator (N2S3) using the MNIST handwritten digit recognition dataset. The first results show better performance of our novel synapse than the traditional nonvolatile artificial synapses.

Keywords

Spiking Neural Network, Memristive Devices, Unsupervised Learning, Digit Recognition

1. INTRODUCTION

Neuromorphic computing has the potential to bring very low power computation to future computer architectures and embedded systems [1]. Indeed parallel neuromorphic computing, by doing computation and storage in the same devices can overcome the Von-Neumann bottleneck. Neuromorphic computing is introduced as an appropriate platform for Big Data analysis and Cloud Computing. Furthermore, many huge projects are running based-on neuromorphic system such as the EU Human Brain Project [2], the DARPA/IBM SYNAPSE project [3] and deep learning research by Google and Facebook among others.

Recently, emerging devices in nano-scale have demonstrated novel properties for making new memories and unconventional processing units. One of those is the memristor that was hypothetically presented by Leon Chua in 1971 [4] and after a few decades, HP was the first to announce the successful memristor fabrication [5]. The unique properties in memristor nano-devices such as, extreme scalability, flexibility because of analog behavior, and ability to remember the last state make the memristor a very promising candidate to apply it as a synapse in Spiking Neural Network (SNN) [6].

In the recent years, there have been several research works using non-volatile resistive nanodevice as a synapse to build a SNN hardware [1, 6, 7]. Forgetting in the biological brain is an important key of adaptive computation, as without forgetting the biological memory soon becomes overwhelmed by the details of every piece of information ever experienced. Consequently, some studies have been done using volatile memory as a synapse in brain-like computing [8, 9, 10]. In this work, we combine both volatile and non-volatile types of artificial synapses. It leads to make a synapse which can forget if the information is not important as well as remember if it is significant data.

Thanks to close collaboration with the nano-electronics research center in the University of Lille (IEMN), we have the opportunity of studying the suitability of different kinds of memristors (TiO₂, NOMFET, magnetoresistive, magnetoelectric) to build a spiking neural network hardware platform. Due to the demonstrated potential of NOMFET (Nanoparticle Organic Memory Field-Effect Transistor) [8, 9] to play the role of a synapse, we use it as a volatile synapse in neuromorphic accelerator. The non-volatile device could be any solid-state memristor. We have chose here the resistive memory presented in [11] as non-volatile memory.

We evaluate the synapse box proposal by comparing it with a single non-volatile memory synapse by simulation on the MNIST handwritten digit recognition benchmark. We use Leaky Integrate and Fire (LIF) neurons in Restricted Boltzmann Machine (RBM) network topology. To run the simulations, we introduce the Neural Network Scalable Spiking Simulator (N2S3), a simulation framework for architecture exploration of neuromorphic circuits.

In the next section we describe the architecture including the neuron and synapse models, as well as the network topology and the unsupervised training algorithm. Then,
2. CIRCUIT DESIGN OF NEURON AND SYNAPSE IN RBM NETWORK

In the biological brain, neurons are computing units. Here we define our computing unit by inspiration from a biological neuron model. The synapse operates as a plastic controller between two neurons. The manufacturability is out of scope of this paper, however in the synapse box, we have applied the real parameters of volatile synapse beside the model of nonvolatile Resistive RAM. As there are varieties of nonvolatile memristor fabrications, finding appropriate one that is compatible with NOMFET seems not too complicated. It worth mentioning that most of the nanodevices have been reported as Resistive RAMs (nonvolatile) as well as NOMFET (volatile) are compatible with CMOS circuits [9].

2.1 Leaky Integrate-and-Fire neurons

The Leaky-Integrate-and-Fire (LIF) neuron model is a well-studied model of neuron. There are three reasons for using LIF in our platform.

- The fabricated model with recent CMOS technology is available [12, 13].
- LIF works effectively in spiking and event-based networks [14].
- LIF models are quite fast to simulate, and particularly attractive for large-scale network simulations [15].

Neurons integrate the spike inputs from other neurons they are connected to. These input spikes change the internal potential of the neuron, it is known as neuron’s membrane potential or state variable. When this membrane potential passes a threshold voltage due to integrated inputs, the action potential occurs, in other words, the neuron fires.

The model is described by the neuron membrane potential:

\[ \tau \frac{dv}{dt} = -v(t) + RI_{syn}(t) \]  (1)

\[ I_{syn}(t) = \sum_j g_{ij} \sum_n \alpha(t - \tau^{(n)}) \]  (2)

where, \( v(t) \) represents the membrane potential at time \( t \), \( \tau = RC \) is the membrane time constant and \( R \) is the membrane resistance. Equation 1 describes a simple parallel resistor-capacitor (RC) circuit where the leakage term is due to the resistor and the integration of \( I_{syn}(t) \) is due to the capacitor. The total input current, \( I_{syn}(t) \), is generated by the activity of pre-synaptic neurons. In fact, each pre-synaptic spike generates a post-synaptic current pulse. The total input current, injected to a neuron is the sum over all current pulses which is calculated in Equation 2. Time \( \tau^{(n)} \) represents the time of the \( n \)th spike of post-synaptic neuron \( j \), and \( g_{ij} \) is the conductance of synaptic efficacy between neuron \( i \) and neuron \( j \). Function \( \alpha(t) = q\delta(t) \), where \( q \) is the injected charge to the artificial synapse and \( \delta(t) \) is the Dirac pulse function. If \( I_{syn}(t) \) is big enough where action potential can pass the threshold voltage, neuron fires. It means there are enough input spikes in a short time window. When there is no or only a few spikes in a time window, the neuron is in the leaky phase and the state variable decreases exponentially. The duration of this time window depends on \( \tau = RC \). The equation is analytically solvable and thus we use the answer of Equation 1 in the network simulation when there is an input spike to improve the simulation performance. In Figure 1a, you can see the Matlab model of a single neuron. When the input voltage passes the threshold, the neuron fires and resets to resting state. The membrane potential stays for an indefinite period, which is called the refractory period, below the reset value.

2.2 Artificial synapse

Before the discovery of the memristor nanodevice, by using state-of-the-art technology, 12 transistors were combined to mimic the behavior of memristor to perform the STDP learning method [16]. Therefore, using a two-terminal and scalable device such as the memristor could save remarkable amount of power and cost specially in modeling large scale Spiking Neural Networks. To model biological synapses, not only do we need a device to be able to store the last activity, but it must also have enough flexibility to achieve Spike Timing-Dependent Plasticity (STDP) for learning. Using memristor as a nonvolatile synaptic memory has been proposed in several works [8, 6, 17, 18]. By using nonvolatile memory, we can guarantee to store the last synaptic weight which is necessary for network training but the synapse can not forget. To be able to have a synapse which is able to forget, scientists used a volatile memory cell [9, 10].

Forgetting is a memory mechanism that helps brain having better functionality. In fact, it is believed that forgetting helps the brain to remember. However, remembering details of many daily activities and information such as shopping list, novel book details or newspapers not only are unnecessary to remember but also might interfere with brain functionality for innovative thinking and data analysis. Basically human brain skips details of insignificant information and remembers the most important, unique and surprising events and information. In neuroscience, memorization is believed to achieve as a result of two types of synaptic plasticity: Short-Term Potentiation (STP) and Long-Term Potentiation [19].
STP is achieved through the temporal changing of a synaptic connection and the decrease to its initial state soon after. In LTP, stimulation iteration causes permanent synaptic weight achievement as it is depicted in Figure 1.b and 1.c. Shorter iteration interval leads to more efficient LTP. By inspiring of this biological theory, we propose a new artificial synapse with the ability to forget insignificant data while storing significant information. The novel synaptic box includes one organic transistor and one resistive RAM.

Resistive RAM is modeled in our previous work [20] and is used here as a nonvolatile memristor in the synapse box. As it is shown in Figure 2 b by changing the doped-undoped regions of device, the conductance will be changed. Bigger doped region leads to more conductivity. Therefore by controlling this boundary between two regions, the conductivity is controlled. The behavior of memristor can be modeled as follows [5]:

\[ v(t) = R_m i(t) \]  
\[ R_m = R_{ON} \frac{w(t)}{D} + R_{OFF} \left( 1 - \frac{w(t)}{D} \right) \]

where \( R_m \) is the variable resistance of memristor, \( w(t) \) is the width of the doped region, \( D \) is the overall thickness of device, \( R_{ON} \) and \( R_{OFF} \) are device resistances while the active region is completely doped (\( w = D \)) and mostly undopped (\( w \rightarrow 0 \)) respectively (Figure 2.b). To model the changing of the conductance, we use the model extracted from Equation 4 and introduced in [11, 21] by considering \( g_{max} = \frac{1}{R_{ON}} \)

and \( g_{min} = \frac{1}{R_{OFF}} \) as the maximum and minimum device conductance respectively.

Organic synaptic memory is a novel memristive device with capability of mimicking synaptic properties especially forgetting ability. Nano-particle Organic Memory Field Effect Transistor (NOMFET) is an organic memristive device made of conjugated molecules and metal nanoparticles (NPs) which is fabricated by the Institute of Electronics, Microelectronics and Nanotechnology (IEMN) at Lille university [8]. We use NOMFET as our volatile device in the synapse box. In the most recent fabrication process [10], NOMFET works at 1 V with a typical response time in the range 100–200 ms. NOMFET is designed particularly for neuro-inspired computing architectures [9]. NOMFET uses charge trapping/detrapping in an array of gold nanoparticles (NPs) with the SiO2/pentacene interface designed to mimic dynamic plasticity of a biological synapse as depicted in Figure 2 [9]. The NOMFET is used as a two-terminal device by connecting drain (D) and gate (G) together and using this terminal as an input. The source (S) is used as output of the device. Equation 5 shows the behavior of NOMFET as a memristor:

\[ i(t) = g(q_{np}(t), v_{ds}(t), t)v_{ds} \]

where \( g \) is the conductance of the device, \( v_{ds}(t) \) is the applied voltage and \( q_{np} \) is the charges trapped in the NP. For more details of physical structure and behavior of NOMFET refer to [9, 10].

Figure 2.c is the synapse box schematic that we apply in our simulation platform to take the advantages of both nonvolatile and volatile artificial synapses. The equivalent circuit of transistor is depicted in Figure 2.d. Actually, weight modification follows the STP rule until reaching the LTP threshold in NOMFET. The modification of nonvolatile device is based on STDP learning. Indeed the NOMFET reacts similar to a high-pass filter (HPF). The stimuli spikes with low frequency are not qualified to pass in forgetting Phase. In LTP, stimuli spikes which have more frequency pass to interfere in learning phase (Figure 1.c).

2.3 Network topology and learning

By using unsupervised learning inspired by biological neural networks, we propose a fully connected network architecture similar to Restricted Boltzmann Machine [22]. To figure out the correlation between the data, STDP helps to adjust the weight if the sensory input spikes are frequent enough to pass the STP and remain in LTP phase. In STDP, if there is output spike in pre-synaptic neuron and shortly after in post-synaptic neuron, the conductance of the synapse between two neurons increases. On the other hand, if the post-synaptic neuron spikes shortly before the pre-synaptic neuron, the conductance of synapse between two neurons decreases. More comprehensive explanation for STDP is beyond the scope of this research, however if readers want to know how plasticity in memristor helps targeting STDP achievement we refer you to [23].

The simulator architecture in our work is event-driven, there is no clock to synchronize the inputs and outputs. Furthermore, by inspiring of biological behavior of brain computing, we apply lateral inhibition to reduce the activity of the neighbors of winner neurons. This method is known as winner-take-all (WTA) strategy [24]. The neuron which reaches the threshold first sends an inhibitory signal to all other neurons in the same layer to reset their states during inhibition time.

The last issue in network architecture that we should address is homeostasis. In STDP learning, the connectivity between two neurons (i.e. the synaptic weight or conductance) is increased when the post-synaptic neuron fires shortly after the presynaptic neuron. This process may be repeated frequently specially with WTA lateral inhibition. Homeostasis is a neuron property that regulates the firing threshold to prevent a neuron to be hyperactive [25]. The idea is to use an adaptive threshold for the membrane potential. If the neuron is too active in a short time window the threshold grows gradually; likewise, when a neuron is not active in a certain time window the threshold is reduced slightly.
3. EXPERIMENTAL VALIDATION

In order to check the effectiveness of the synapse, we propose a spiking neural network simulator. Our requirements for the simulator are: speed (thus event-driven simulation and concurrency), scalability (thus high-level abstraction and distributability), and adaptability (possibility to model different synapses, soma, and network topology). After the presentation of the simulator, we describe our experimental setting and discuss the simulation results that show an improvement in recognition rate for the synapse box with respect to the simple nonvolatile synapse.

3.1 N2S3 (Neural Network Scalable Spiking Simulator)

The most popular neural network simulators in the neuroscience community such as Neuron [26], Brian [27] or NEST [28] can provide different levels of abstraction. However, they are clock-driven and the model of memristor as a synapse is not considered. Xnet [29] is an event-driven simulator but its capabilities do not address some of our requirements such as scalability and concurrency. Furthermore, XNet is not available to us. We have thus developed a new simulator dedicated to the design of nano electronic spiking neural networks.

N2S3 (Neural Network Scalable Spiking Simulator, pronounced “Nessy”) is an event-driven simulator: it does not iteratively solve differential equations but uses the analytical solution of these equations and updates the states of the neurons only upon the arrival of input spikes. To address our concurrency and distributability requirements (ability to scale out a simulation on several computers to handle large networks) we have chosen to use the Scala programming language [30] along with the Akka actor library [31]. The internals of N2S3 are thus based on the exchange of messages between actors, mimicking the exchange of spikes between neurons. N2S3 has been developed from the ground up for extensibility, allowing to model various kinds of neuron and synapse models, various network topologies (it is not restricted to Boltzmann machines), various learning procedures, various reporting facilities, and to be user friendly with a domain specific language to express the experiments the user wants to simulate. It will be available as open source software before the end of 2016 at https://sourcesup.renater.fr/projects/n2s3.

At the moment, N2S3 can read AER files and the MNIST files and convert them to spikes that are sent to an artificial neural network of any topology (we currently have shortcuts to model fully connected multilayer networks such as Restricted Boltzmann Machines and are working on deep belief network topologies). The spikes are carried by messages between actors representing some subsets of the neurons of the network, and are fully or partially synchronized to offer a tradeoff between accuracy and concurrency. In the experiment below, messages are fully synchronized, and thus ordered by timestamps so that the accuracy of the simulation is maximized. A full description of N2S3 is out of the scope of this paper and will be the object of future article.

3.2 MNIST recognition improvement

We have used the MNIST training dataset of handwritten digits [32] to train and test the performance of neural networks based on the synapse box. The training set consists of 60000 digits between 0 and 9 and each handwritten number is a 28 × 28 pixel image. In this simulation, we present the full dataset (60000 images) and full images. Each pixel is connected to one input buffer neuron. Pixel intensity is between 0 to 255 and is transferred to 0 to 22 Hertz spiking frequency using a Poisson distribution during a 350 ms presentation window. Based on previous similar work [33], we have chosen a delay of 150 ms between two images. Therefore, there is sufficient time for membrane potentials of all neurons to reset back to initial values. The network connection weights are between 0 and 1 initialized using a Gaussian distribution.

The hardware platform is a 4 core i7-3687U CPU (2.10GHz × 4). We have simulated different network topologies consisting of 2 fully interconnected layers, with a fixed input neuron number (28 × 28 = 784) and different output neuron number. The neuron model is LIF and we evaluate two types of synapses: non-volatile (NV) and proposed synapse box (volatile/nonvolatile or VNV).

To measure and evaluate the network classification accuracy after a fully unsupervised learning period consisting of the presentation of the full MNIST data set, we label the output neurons using 10000 samples of MNIST: After training, we stop all synaptic modification processes such as STDP, STP and LTP. We assign a number class to the output neuron which has most frequent firing rate during the presentation of the 10000 labelling samples. Using these labels, we then evaluate the recognition rate of the network on 10000 different test samples by comparing the known class of the input with the class of the most firing output neuron.

As it was observed in similar works [33] and [11], the recognition rate depends on the number of neurons and synapses, and the number of repetitions of the presentation of the dataset to the network. In the experiments, we present 60000 digit the dataset the smallest number of time that is necessary for learning. That is 1 time for less or equal to 50 output neurons, 2 times for less or equal to 100 output
neurons and 3 times for more output neurons. With these numbers of presentations, we obtain recognition rates that are comparable to the state of the art. The running time of the simulations are also comparable to those of similar experiments though it is difficult to make accurate comparisons.

An example of the conductance weights learned in N2S3 on the MNIST dataset is shown in Figure 3 for 100 output neurons. As it is obvious in the figure, the border of each digit did not pass the NOMFET high-pass filter because of low frequency changes. This is the impact of forgetting properties in synapse box to skip unimportant data or noise. To demonstrate the functionality of the synapse box, we compare the recognition rate of networks of the same topology but using different synapse models: a simple nonvolatile synapse (NV) model and the volatile-nonvolatile (VNV) synapse box model. We have run 10 simulations for each number of output neuron and each synapse. The results are summarized in Figure 4 showing the distribution of the recognition rates for each configuration. We can conclude that using the synapse box improves the recognition rate in average by a small but consistent margin.

Although it is not shown in Figure 4, we have also made the comparison using 300 output neuron and the best recognition rate we have obtained is 89.4%.

4. CONCLUSION
In this study, we have introduced a novel synapse box with the possibility to forget and remember inspired from biological synapse properties. This synapse box is composed of a volatile memristor (NOMFET) followed by a nonvolatile resistive RAM. The volatile memristor acts like a high-pass filter to enhance short term plasticity and the nonvolatile resistive RAM enables long term potentiation. Both work together in the spike timing dependent plasticity unsupervised learning process.

In addition, in this work we have also announced a new event-based simulator, N2S3 (Neural Network Scalable Spiking Simulator). It is specifically designed to simulate hardware spiking neural networks. N2S3 is quite flexible to explore different network architectures, synapse and neuron models to help design hardware architectures and VLSI circuits. To evaluate and verify the new synapse box as well as the functionality of the simulator, we have used the MNIST handwritten digit dataset. The first results demonstrate an improvement in recognition rate by using the synapse box over a single nonvolatile memristor synapse. We will continue to explore the various parameters and device combinations to help design the most efficient hardware neural networks as possible.

For future works, we also propose to study different neural network topologies such as deep belief, recurrent and convolutional neural networks to evaluate the synapse box benefits and costs (area, energy, manufacturability, variability) in other contexts.

5. ACKNOWLEDGEMENTS
This work has been partly funded by IRCICA (Univ. Lille, CNRS, USR 3380 - IRCICA, F-59000 Lille, France).
We would like to thank Dr Fabien Alibart for his support using nanodevices and Dr Philippe Devienne for his strong support when writing this paper.

6. REFERENCES


