Long Short-Term Memory with Spin-Based Binary and Non-Binary Neurons

Shadi Sheikhfaal, Meghana Reddy Vangala, Adedoyin Adepegba, and Ronald F. DeMara

Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL
shadi@knights.ucf.edu, meghanareddy016@knights.ucf.edu, adedoyin.a@knights.ucf.edu, and ronald.demara@ucf.edu

***Abstract*—In this paper, we develop a low-power and area-efficient hardware implementation for Long Short-Term Memory (LSTM) networks as a type of Recurrent Neural Network (RNN). The LSTM network herein employs Resistive Random-Access Memory (ReRAM) based synapses along with spin-based non-binary neurons to achieve energy-efficiency while maintaining comparable accuracy. The proposed neuron provides a novel activation mechanism with five levels of output accuracy to mimic the ideal tanh and sigmoid activation functions. We have examined the performance of an LSTM network for name prediction purposes utilizing ideal, binary, and the proposed non-binary neuron. The comparison of the results shows that our proposed neuron can achieve up to 85% accuracy and perplexity of 1.56, which attains performance similar to algorithmic expectations of near-ideal neurons. The simulations show that our proposed neuron achieves up to 34-fold improvement in energy efficiency and 2-fold area reduction compared to the CMOS-based non-binary designs.**

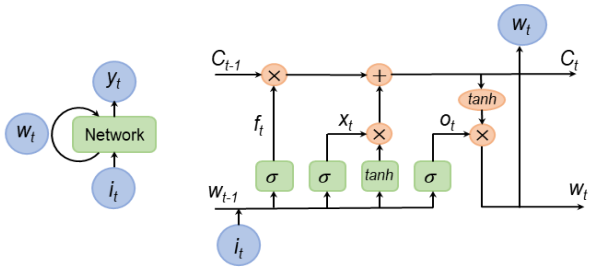
Keywords—Long Short-Term Memory, Binary Stochastic Neuron, Activation functions, Non-Binary Neuron

I. Introduction

Long Short-Term Memory (LSTM) networks, as a form of Recurrent Neural Networks (RNNs), have achieved noticeable recognition due to their ability to process sequential data and gathering the impacts of the input data over time. LSTMs have demonstrated high performance in various sequence prediction problems in applications such as speech recognition and machine translation. Fig. 1 shows the basic RNN and LSTM structures. Both networks have a feedback loop in their recurrent layer to sustain the information over time. LSTM utilizes additional units including a memory cell capable of storing information for long periods [1].

There are several research works exploring hardware implementation of RNNs which employ a non-von-Neumann architecture, based on the Compute-in-Memory (CiM) designs to provide highly parallel and efficient models [2, 3]. Most of the previous designs utilize emerging Non-Volatile Memory (NVM) devices such as Resistive Random-Access Memory (ReRAM) [4], to implement the Multiplication and Accumulation (MAC) operation via the intrinsic weighted summation capability of cross-bar designs based on CiM architecture. However, these designs require significant power and area due to the employment of CMOS-based non-linear sigmoid and tanh neurons, as their main thresholding functions. The utilized CMOS-based neurons in prior works [2, 3] require large built-in truth tables with extra clock cycles that lead to higher area and energy consumption.

In this paper, we implement an LSTM network with ReRAM-based synaptic crossbar arrays and spin-based non-binary neurons mimicking the ideal sigmoid and tanh thresholding functions while maintaining accuracy. To achieve an area and energy-efficient neuron design, we employ probabilistic spin logic devices referred to in the literature as probabilistic bits (p-bits) [6] to develop a shared neuron unit in the hidden layer. We investigate the performance and efficiency of the LSTM network with three distinct neuron structures.



(a) (b)

Fig. 1. (a) Basic RNN structure, (b) LSTM

II. LSTM Structure

Fig. 1. (a) shows the basic RNN structure. The RNN output depends on both the current sample (i_t) and the previously calculated network state (w_t) as the network input. Unlike ANN, RNN has a feedback loop which enables it to store the previous states and make the future decision based on the previous values. The LSTM network is designed to overcome the problem of vanishing gradients that occurs while using the backpropagation technique in RNNs [5]. Fig. 1. (b) indicates an LSTM cell that contains input gate x_t, forget gate f_t, and output gate o_t. The forget gate decides which information from the previous cell state must be preserved or forgotten. This decision is taken using a sigmoid layer with an output between 0 and 1 [6]. The input gate decides what values of the new cell must be written to the cell state. The sigmoid layer determines the input values (a concatenation of new input values and output values from previous states), while the tanh layer produces a vector of new candidate values. The output gate works based on given inputs and previous state values. The output vector is obtained by multiplying a new cell state which is normalized to values between -1 to 1 using tanh activation function and output of sigmoid layer that decides the output. The dimensions of all the gates are the same as the dimensions of the hidden state [7]. The computational equations of LSTM are given below:

$$x = \sigma(i_t U^x + w_{t-1} W^x + b_x) \quad (1)$$

$$f = \sigma(i_t U^f + w_{t-1} W^f + b_f) \quad (2)$$

$$o = \sigma(i_t U^o + w_{t-1} W^o + b_o) \quad (3)$$

$$g = tanh(i_t U^g + w_{t-1} W^g + b_g) \quad (4)$$

$$c_t = c_{t-1} \odot f + g \odot x \quad (5)$$

$$w_t = tanh(c_t) \odot o. \quad (6)$$

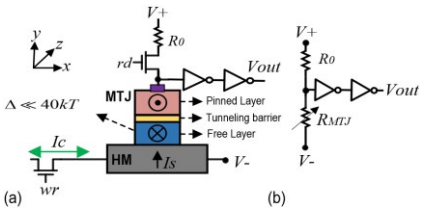


Fig. 2. (a) The building block of non-binary neuron (p-bit), and (b) the equivalent read circuit [8].

Three main operation types can be observed from the above equations: nonlinear functions (sigmoid $\boldsymbol{\sigma}$ and hyperbolic tangent $tanh$), matrix-vector multiplication (e.g., $w_{t-1}W^x$ and i_tU^x), and element-wise multiplication (e.g., $g \odot x$).

III. Neuron Design

A. Probabilistic Spintronics

The spintronic device used herein is derived from Magnetic Tunnel Junction (MTJ) which has a probabilistic behavior with a design of 1-Transistor-with-1- MTJ structure called an embedded probabilistic bit (p-bit) [8]. Due to the very low energy barrier of the free layer, the p-bit stochastically switches between its Parallel (P) and Anti-Parallel (AP) states. The mean retention time for an MTJ (τ) is given by (7).

$$\tau = \tau_0 exp(\Delta/kT) \quad (7)$$

where k is Boltzmann's constant, τ_0 is a material-dependent parameter called the attempt time, and T is temperature [8].

Fig. 2(a) shows the structure of the p-bit device which consists of a Spin Hall Effect Magnetic Tunnel Junction (SHE-MTJ) with a circular unstable (low energy barrier) nanomagnet (Δ≪40kT) [8], to which two CMOS inverters are connected to amplify the output. The MTJ in the device is unstable with two ferromagnetic layers as a *pinned* layer and the *free* layer, separated by a thin oxide barrier on top of a Heavy Metal (HM) nanowire [9]. The *pinned* layer has a fixed orientation while the free layer can be oriented as *parallel (P)* and *antiparallel (AP)*. As shown in Fig. 2(a), the charge current (I_c) injected to HM in the +x (/−x) direction affects the resistance levels [10]. This charge current will produce a spin current (I_s) and a Spin-Orbit Torque (SOT) in +y (/−y) direction as oppositely directed spin vectors are accumulated on each surface of the HM. The direction of the charge current affects the spin current which further changes the magnetization configuration of the free layer in the ±z direction [11]. Sigmoidal function can be derived by taking a long-time average of magnetization fluctuations of the low energy barrier nanomagnet driven by spin-current, as shown in Fig 3(a). In Fig. 2(b) an equivalent read circuit of a SHE-MTJ based p-bit is given in which reading operation is done by sending a small read voltage to MTJ terminals ($V+$ and $V-$) to sense its resistance (R_{MTJ}). The R_{MTJ} and the reference resistor R0 are then used to construct a resistive voltage divider, with the reference resistor being assigned to the MTJ average conductance ($R_0^-1 = \mathrm{GP} + \mathrm{GAP}/2$) where GAP and GP are the AP and P state conductance. The voltage from the voltage divider is given as input to the CMOS inverters and its output voltage (V_{out}) will stochastically fluctuate between "0" and "1", and the probability of each value is controlled by the input charge current. Thus, a p-bit device generates a stochastic output which is analogous to the output of a sigmoid activation function whose steady-state probability is modulated by an input current. For example, if the input current is a large positive number, the stochastic output of this device will be "0" with a high probability. However, if there is no input current, the output will randomly fluctuate between "0" and "1" with an equal probability of 0.5.

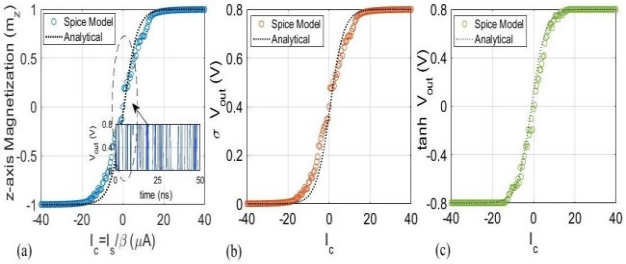


Fig. 3. Time-averaged behavior of SHE-MTJ based p-bit device. (a) magnetazitaion fluctuation. (b) and (c) are the implemented sigmoid and tanh behaviors respectively.

B. Binary and Non-Binary Neurons

As discussed in the previous sections, LSTM networks require sigmoid and tanh-based neurons for multiple gating purposes. The current-controlled p-bit device shows an analogous behavior to the sigmoid function in an average time interval. Fig. 2. (a) shows circuit implementation of the p-bit device. A sigmoidal behavior can be achieved by connecting an inverter to VDD and GND. In Fig. 3. (b), the sigmoid function output is indicated by the black dotted curve, and the p-bit output average is indicated by the red-circle curve which is almost the same as the sigmoid output curve. In the same way, the nonlinear hyperbolic tangent or tanh function can be designed using a sigmoid function as $\tanh(x) = 2\sigma(2x) - 1$ whose output values are between "+1" and "-1". This can be achieved by connecting an inverter to VDD and -VDD in the p-bit device. In Fig. 3(c), the tanh function output is indicated by the black dotted curve, the green-circle curve indicates the time-averaged output of the modified p-bit device (*tanh (I_c)*). This figure shows that the output of p-bit at each time step depends on the input, a zero input gives an output of either "-1" or "+1" with equal probability, a positive input I_c gives a high probability to output a positive value, and vice versa.

Therefore, the time-averaged output of the p-bit device can provide both sigmoid and tanh function behaviors via slightly different circuit designs. However, for practical implementation, the p-bit device gives a binary output of either "0" or "1" at a given time. Conversely, ideal sigmoid and tanh functions do not have a limited binary state as outputs but vary within a limited range based on the input value. To utilize a p-bit device as a practical activation function to achieve higher accuracy levels there is a need for a novel complementary activation circuit and mechanism. In any p-bit device, the stochasticity is highest for input current values that are near to zero and the stochasticity reduces as the input current values reach their highest or lowest levels. This behavior of a p-bit can be used to implement a non-binary neuron. This behavior is extracted in the proposed design by running the p-bit device multiple times for the same input obtaining a symmetric range of output voltages.

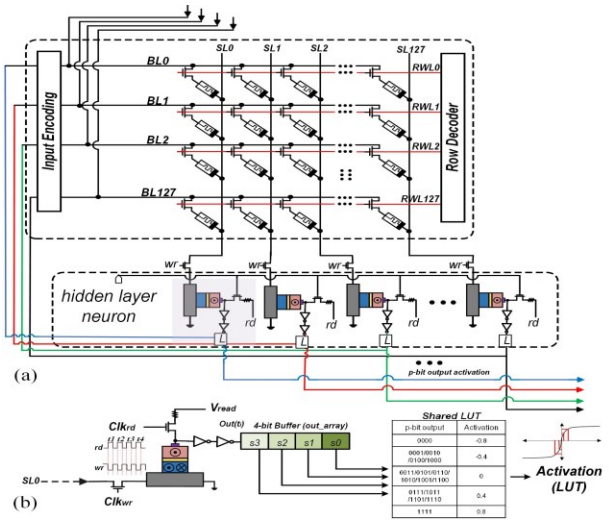


Fig. 4. The proposed spin-based LSTM network with non-binary neurons.

These output voltages are stored and mapped to a low-overhead Look-Up Table (LUT) which contains the voltage values. This technique maintains the low-power and low-area properties of p-bit along and utilizes it to achieve an enhanced non-binary state. We improved the p-bit based stochastic neuron for hardware implementation by adding two components, as shown in Fig. 4. To latch the output voltage of the p-bit circuit, a 4-bit buffer is inserted first corresponding to the four times of applying the crossbar output to the p-bit device. Second, the neuron output is formed using a LUT. We synchronized the p-bit device's write/read access transistors in order to prevent multiple crossbar computations. This method allows the design to keep a reliable crossbar output current and applies it to the neuron unit as needed. As shown in Fig. 4, we consider two complementary signals for wr and rd. The wr signal goes high for each sample and based on the crossbar output current the p-bit device is programmed. To read out the p-bit resistance and produce the output bit, the wr signal goes low and the rd signal goes high. The 4-bit buffered data is then given to the converter LUT which is prestored with the sampled floating-point activation values corresponding to output combinations in the buffer. For example, if the buffer content is 001, the LUT selects -0.4 as the output. This value can be triggered by either 0001/0010/0100/1000 p-bit output bitstreams. Such non-binary neuron design is applicable in a variety of ANN applications needing non-linear and deterministic tanh and sigmoid activation functions.

IV. Results

We evaluate the proposed LSTM design performance starting with device-level modeling of memristive synapse and p-bit based neuron components. We utilized the SPICE model for memristors with the Ag-Si memristor device parameters from [12]. The SHE-MTJ model is developed in Verilog-A, incorporating the Landau Lifshitz–Gilbert (LLG) equation to model the free layer magnetization dynamics and nonequilibrium Green's function (NEGF) to estimate the resistance range (RP, RAP). We then combine the SPICE models of CMOS transistors and memristors under the 14nm PTM-MG library [13]. At the circuit level, we developed crossbar arrays under two sizes (32×32, 128×128) with p-bit neurons in HSPICE. We implemented all peripheral circuits including row address decoders, array controller, etc. in Synopsys Design Compiler. As application-level analysis, we built three distinct name predictor LSTM networks via ideal, binary, and the proposed non-binary neuron, employing the popular names dataset available as national data [14].

Table I: The comparison of proposed non-binary neuron with CMOS-based designs.

	32x32			128x128		
xbar Size	[12]	[13]	Here	[12]	[13]	Here
xbar #	68	68	68	5	5	5
Area (mm^2)	0.17	0.07	0.06	0.06	0.02	0.02
Energy (uJ)	N/A	4.04	0.14	N/A	1.03	0.03

A. Circuit-Level Analysis

Figure 5 shows the SPICE simulation waveforms of the p-bit based non-binary neuron, verifying its functionality. Here we evaluate the neuron output two times (p-bit 1 to p-bit 2) under five input currents for four clock cycles. Here, I_{sum} denotes the weighted summation of input currents realized by the resistive sub-array, ranging from -50μA to +50μA, flowing into the p-bit device. When the I_{sum} is -50μA or +50μA, the output of both p-bit devices for the entire four clock cycles are "1" and "0", respectively, indicating the deterministic behavior of the neuron based on these charge currents. These outputs will later denote 0.8V and -0.8V, via the LUT. When the I_{sum} is -5μA, we observe different outputs for each p-bit device. However, both outputs will later be mapped to a shared value (-0.4V).

Additionally, we compared the area and energy consumption of the proposed design with [15] and [16] CMOS-based designs, under two distinct sub-array sizes as tabulated in Table I. The simulations show that our proposed neuron achieves up to 34× improvement in energy efficiency and 2× area reduction compared to the CMOS-based non-binary designs. The energy consumption results for [15] could not be appropriately reported.

B. Experimental Results

Figure 6 shows the experimental results for three distinct neuron designs including loss, perplexity, and accuracy

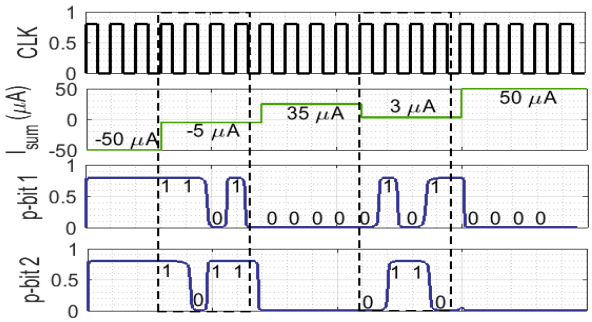


Fig. 5. The transient simulation result of the neuron based on the crossbar SL current.

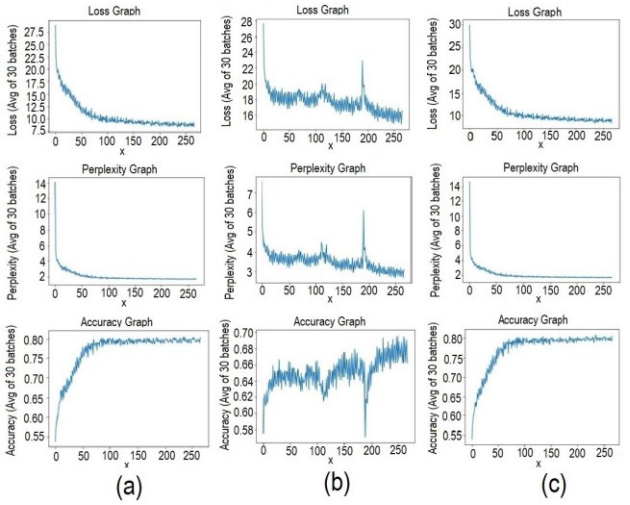


Fig. 6. The experimental results of the LSTM network with (a) ideal, (b) binary and (c) proposed non-binary neuron.

fluctuations for all cases. Unlike accuracy, for loss and perplexity parameters, lower values are preferred. The plotted data is an average of 30 training sample batch sets. The accuracy indicates the performance of the neural network while the perplexity graph evaluates the currently implemented network regarding the sample data modeling. In Fig. 6. (a), the ideal sigmoid neuron displays the limits on possibility with an approximation for all plots. In the binary neuron case shown in Fig. 6. (b), there is a sharp rise in accuracy in the first sets of batches. However, it initially does not reach the performance of the ideal sigmoidal model (Fig. 6. (a)). Consequently, the results of the binary case have a long tail that starts around set number 50, in which the system gradually improves as it progresses towards the end of the batches. Additionally, the perplexity graph shows that disturbance from discontinuity of the binary activation causes the training algorithm to struggle in modeling the samples using the network. After 8,000 training samples, the network with the binary neuron shows 58% degradation at modeling the data compared to the ideal sigmoid neuron.

Utilizing the proposed non-binary neuron, the results are very close to the ideal case as shown in Fig. 6. (c). The enhanced activation mechanism allows it to mimic the ideal sigmoidal system. This is reflected in the perplexity graphs converging to similar values, with the proposed non-binary neuron with only 7% degradation compared to the sigmoidal system. However, the proposed neuron, also starts with a slightly slower training speed, as the binary activation function. But this tail is much shorter, lasting over the course of approximately 1,050 training samples.

V. Conclusion

Hardware implementation of an ideal low-power neuron with a small area overhead is a key research challenge for ANNs. In this paper, we developed energy and area-efficient hardware implementation for LSTM networks via novel spin-based non-binary neurons. The LSTM network herein employs ReRAM based crossbar arrays to achieve energy-efficiency while maintaining comparable accuracy. The proposed neuron provides a novel activation mechanism mimicking the ideal tanh and sigmoid activation functions. The performance evaluation of an LSTM network for name prediction purposes utilizing ideal, binary, and the proposed non-binary neuron shows that the proposed neuron can achieve up to 85% accuracy and perplexity of 1.56, similar to algorithmic expectations of near-ideal neurons. The circuit-level simulations show that our proposed neuron achieves up to 34× improvement in energy efficiency and 2× area reduction compared to the CMOS-based non-binary designs.

Acknowledgment

This work was supported in part by the Center for Probabilistic Spin Logic for Low-Energy Boolean and Non-Boolean Computing (CAPSL), one of the Nanoelectronic Computing Research (nCORE) Centers as task 2759.006, a Semiconductor Research Corporation (SRC) program sponsored by the NSF through CCF 1739635.

References

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[2] Y. Long, et al, "Reram crossbar based recurrent neural network for human activity detection," in *2016 IJCNN*, 2016, pp. 939-946: IEEE.

[3] Y. Long, et al, "ReRAM-based processing-in-memory architecture for recurrent neural network acceleration," *IEEE Transactions on VLSI,* vol. 26, no. 12, pp. 2781-2794, 2018.

[4] H.-S. P. Wong *et al.*, "Metal–oxide RRAM," *Proceedings of the IEEE,* vol. 100, no. 6, pp. 1951-1970, 2012.

[5] K. Greff, et al, "LSTM: A search space odyssey," *IEEE transactions on neural networks,* vol. 28, no. 10, pp. 2222-2232, 2016.

[6] K. Smagulova, et al, "A memristor-based long short term memory circuit," *Analog Integrated Circuits Signal Processing,* vol. 95, no. 3, pp. 467-472, 2018.

[7] X. Zhu, et al, "Long short-term memory over recursive structures," in *International Conference on Machine Learning*, 2015, pp. 1604-1612: PMLR.

[8] K. Y. Camsari, et al, "Stochastic p-bits for invertible logic," *Physical Review X,* vol. 7, no. 3, p. 031014, 2017.

[9] L. Liu, et al, "Spin-torque ferromagnetic resonance induced by the spin Hall effect," *Physical review letters,* vol. 106, no. 3, p. 036601, 2011.

[10] S. Sheikhfaal and R. F. Demara, "Short-Term Long-Term Compute-in-Memory Architecture: A Hybrid Spin/CMOS Approach Supporting Intrinsic Consolidation," *IEEE Journal on Exploratory Solid-State Computational Devices Circuits,* vol. 6, no. 1, pp. 62-70, 2020.

[11] A. Roohi, R. Zand, D. Fan, and R. F. DeMara, "Voltage-based concatenatable full adder using spin hall effect switching," *TCAD,* vol. 36, no. 12, pp. 2134-2138, 2017.

[12] L. Gao, et al, "Analog-input analog-weight dot-product operation with Ag/a-Si/Pt memristive devices," in *2012 IEEE/IFIP 20th VLSI-SoC*, 2012, pp. 88-93: IEEE.

[13] *PTM*. Available: http://ptm.asu.edu/

[14] *Beyond the Top 1000 Names* Available: https://www.ssa.gov/oact/babynames/limits.html

[15] M. N. Bojnordi and E. Ipek, "Memristive boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning," in *2016 IEEE HPCA*, 2016, pp. 1-13: IEEE.

[16] A. Ardakani, et al, "VLSI implementation of deep neural network using integral stochastic computing," *IEEE TVLSI,* vol. 25, no. 10, pp. 2688-2699, 2017.