Stock Price Prediction Based on an Energy-Efficient Spiking-LSTM Hardware Accelerator

Yuancong Wu\(^1\)*, Y Liu\(^1\) and T P Chen\(^2\)

\(^1\) State Key Laboratory of Electronic Thin Films and Integrated Devices, University of Electronic Science and Technology of China, Chengdu 610054, P. R. China

\(^2\) School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

*Email: wycong@std.uestc.edu.cn

Abstract. Inspired by the way the human brain thinks, the neuromorphic system applies the principles of biological brains to computer architecture, providing low-energy, distributed, and massively parallel advantages for brain-inspired systems. This work presents an energy-efficient spiking long short-term memory (sLSTM) neural network hardware accelerator for sequence prediction applications, containing 256 neurons and 64k synapses in 0.96 mm\(^2\) area. The sLSTM model can process time-dependent data and realize long-term and short-term memory to forget, memorize selectively. A leaky integrate and fire (LIF) neuron model is proposed to characterize the stimulation of neuronal membrane potentials using simple digital logic circuit without any multipliers, which extremely reduces the power consumption of the hardware system. Accordingly, the chip achieved an energy efficiency of 10.3 uJ@50 MHz per sample and a predicting accuracy of about 93.2% in sLSTM neural network model using the stock price of Google from Yahoo finance, based on the modified LIF neuron.

1. Introduction

In recent years, artificial neural networks (ANNs) represented by deep learning have achieved excellent performance in games [1,2], image classification [3-5], signal processing [6,7], and natural language processing [8,9], etc. Inspired by the human brain hierarchy and synaptic frameworks, ANNs connect trillions of synapses to simulate brain behavior. The computationally intensive nature of artificial neural networks has led to strict performance and energy bottlenecks on traditional von Neumann computing platforms. Whereas spiking neural networks (SNNs) [10,11], which have the sparsity of the synaptic spiking inputs and the asynchronously event-driven nature of neurons, can be leveraged by energy-efficient hardware implementations and offer significant energy reductions as compared to conventional artificial neural networks.

On the one hand, by using the reconfigurable digital logic platform, ref. [12] proposed an event-based multi-convolution system based on the field-programmable gate array (FPGA). Hamid et al. developed a reconfigurable and efficient 2-dimension neuron model capable of extending to higher dimensions [13]. On the other hand, based on the application-specific integrated circuits (ASIC), IBM has developed an adaptive, scalable plastic electronic neuromorphic system chip, TrueNorth [14]. Intel designed Loihi [15], an autonomous learning neuromorphic chip with 130,000 neurons and 130 million synapses. At present, a large amount of research works [16-18] are proposed to map deep neural networks on neuromorphic hardware.
However, these abovementioned feed-forward neural networks cannot process time-dependent data. Conventional recurrent neural network (RNN) is difficult to learn long-term dependencies. The long short-term memory (LSTM) model improves the RNN through a gating mechanism that allows long-term and short-term memory to forget, memorize selectively, and output information, enabling it to capture long-term time-dependent relationships effectively. Therefore, LSTM has become a prominent and successful model for time-series processing and has achieved success in machine translation, sentiment analysis, automatic question answering, and sequence detection [19, 20] in recent years.

In this article, we propose a spiking neuromorphic core, and extend the application of deep learning recurrent neural networks to the field of neuron morphology, transforming it into spiking-LSTM neural networks. To promise accuracy, speed, and cost, the LIF neuron [21] model is implemented using simple arithmetic operations such as addition, subtraction, and shift instead of the physical multipliers. This chip occupies only 0.96 mm² in the 55nm process and consumes 10.3 μJ energy-efficient per sample at 50 MHz clock frequency. A sLSTM neural network is mapped on this chip, which achieves a prediction accuracy of 93.2% using the stock price of Google from Yahoo.

2. Architecture
In this work, a Xilinx Spartan-6 FPGA central platform is used to control the neuromorphic chip and as data interface, as shown in figure 1. The operations are distributed from the FPGA to the neuromorphic chip by the two asynchronous first-in-first-out (FIFO). The neuron output, in the form of the first-spike-time coding, is decoded by the spike decoder and sent to the central platform for other post-processing. In the FPGA platform, the entire system is scheduled through the cooperation of an embedded Reduced Instruction-Set Computer (RISC) processor soft-core MicroBLaze, PLBv46 bus, and other peripheral devices interfaces.

The proposed chip contains 256 digital LIF neurons, 64k synapses, communication interface, data buffer, spike codec, and neuromorphic controller. Each neuron consists of a LIF model and an arithmetic logic unit (ALU) to support spiking-LSTM model mapping and other arithmetic operations. As shown in figure 2, neurons are connected in crossbar, and can accommodate up to 64k synapses. A reconfigurable synaptic mesh structure is proposed through synapse activation. Activating synapses in the crossbar represents densely connections between inputs and neurons, or sparse connections. The weights of the synapse are stored on the axons instead of the synaptic crossbar. The spikes generated by the output neurons are either returned to the input via the spike synchronization unit or output directly.
3. Spiking-LSTM Module and Hardware Implementation

A spiking-LSTM neural network is introduced for predicting the future sequential data. The single sLSTM, as shown in figure 3, has input, forget and output gates, activated by sigmoid or hyperbolic tangent. The input gate determines the information retained by the previous hidden state and the value to be updated in the current input. The forget gate decides what information should be thrown away or kept. Information from the previous hidden state and information from the current input is passed through the sigmoid function. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep. The output gate determines what information needs to be carried in the hidden state obtained by multiplying the tanh output and the sigmoid output. The concatenated information of the previous time and the current time is multiplied by a weight matrix to perform a linear transformation. The data processing methods of input gate, forget gate and output gate are the same, except that the weight matrix is different.
Figure 3. Block diagram of a single sLSTM model.

Its output $h_t$ is generated based on the equations shown in equation (5):

\[
\begin{align*}
    f_t &= \text{sigmoid}(W_f \odot [h_{t-1}, x_t] + B_f) \\
    i_t &= \text{sigmoid}(W_i \odot [h_{t-1}, x_t] + B_i) \\
    o_t &= \text{sigmoid}(W_o \odot [h_{t-1}, x_t] + B_o) \\
    c_t &= f_t \ast c_{t-1} + i_t \ast \tanh(W_c \odot [h_{t-1}, x_t] + B_c) \\
    h_t &= o_t \ast \tanh(c_t)
\end{align*}
\]  

(1) (2) (3) (4) (5)

$f_t$, $i_t$, and $o_t$ are the forget gate, input gate and output gate vectors, respectively; $c_t$ and $h_t$ are the cell state and output vectors, respectively; and $W_f$, $W_i$, $W_o$, and $W_c$ are trained weight matrices.

Figure 4. The time slice division management mechanism in the proposed chip.

A special neural hardware based on an event-driven was proposed in which the input vector $x_t$ in time window $t$ and output vector $h_{t-1}$ in time window $t-1$ of an sLSTM requires synchronization. As shown in figure 4, a mechanism of time slice division management is adopted, and a spiking-LSTM cell
operation is performed once in each time window \( t \). The obtained output vector \( h_{t-1} \) is synchronized to the start of the next time window by the spiking synchronization module, which consists of a spiking buffer and a clock counter. \( h_{t-1} \) is combined with the input vector \( x_t \) and sent to the sLSTM for the calculation of the current time window. The input and output vectors are both adapted to the first-spike-time coding format based on a cycle of 256 clock steps. The value represented by the spike generated at the \( n \)th time step is \( n - 128 \). This means that the input and output value are in the range of \(-127\) to \(128\).

As shown in figure 5, the structure of a single LIF neuron model, supporting integration, leak, fire, and reset operations. The membrane state follows the dynamics:

\[
V_i(t) = Y_i \cdot \left[ V_i(t-1) + \sum_{j=1}^{256} C_{ij} \cdot W_{ij} \cdot S_{in}^j(t-1) - \alpha t \right] + (Y_i - 1) \cdot V_{th}^i \\
\text{if } V_i(t) > V_{th}^i: S_{out}(t) \leftarrow 1, V_i(t) = -V_{rst}^i
\]

(6)

\( S_{in}^j(t) \) is the input spike on axon \( j \) at time \( t \); \( W_{ij} \) is the synaptic weight; \( C_{ij} \in \{0,1\} \) is the synaptic connectivity between axon \( j \) and neuron \( i \); \( \alpha_t \) is the discharge coefficient of the membrane potential over time; \( Y_i \in \{0,1\} \) is the neuron \( i \) reset coefficient; \( V_i(t) \) is the membrane potential of neuron \( i \) at time \( t \); \( V_{th}^i \) is the fire threshold of neuron \( i \); \( V_{rst}^i \) is the membrane potential of the reset state of the neuron \( i \); and \( S_{out}^i(t) \) is the output spike on neuron \( i \).

An energy-efficient hardware implementation of the LIF neuron model is proposed in this work. The hardware design of a LIF neuron is shown in figure 6, which consists of 256 accumulator cells, a 4-level adder-tree, a membrane potential register, a neuron output register, a discharge coefficient register, an adder, a subtractor and a multiplexer. The membrane potential in the proposed modified LIF neuron is completely realized by a simple digital logic circuit without any multipliers, which extremely reduces the power consumption of the hardware system.
4. Methods

This study proposed a spiking-LSTM neural network model for stock price trend prediction and mapped the model into the proposed sLSTM hardware accelerator in the form of spike transformation. The model consists of three parts: input layer, hidden layer and output layer. The input layer preprocesses the input data to meet the input requirements of the model. The hidden layer is composed of two layers of sLSTM cells for timing feature extraction. The sLSTM cell is affected by the input data at the current time and the previous time. The output layer weights the feature vectors extracted by the hidden layer to obtain the prediction result. The proposed framework is illustrated in figure 7. The model selects the stock price data of Google from Yahoo finance as the dataset. The model splits the dataset into training set, test set and validation set according to the ratio of 7:2:1, and respectively adopts adam optimizer, momentum-based optimizer and stochastic gradient descent (SGD) optimizer for training. The training and prediction results of the model are shown in figure 8, and a predicting accuracy of about 93.2% is achieved.
5. Results

The proposed chip is manufactured by 55nm CMOS process, containing 256 neurons and 64k synapses. Figure 9 shows a photograph of this chip with a 0.96 mm² area, including IO PADS. The core supply voltage of the chip is 1.2 V, and the IO signal level is 3.3 V. The spiking-LSTM neural network has been implemented on the chip with a maximum operating frequency of the chip is 400 MHz. Also, the latency to predict output in this hardware is approximately 4.82 us. Under 1.2 V supply voltage and 50 MHz working frequency, the operating power consumption of the chip is 12 mW. It consumes about 10.3 μJ at 1.2 V per sample and is up to 1.6x, 128x, and 636x energy efficient than NIVIDA Tegra X1, NIVIDA Titan X, and IBM TrueNorth. A summary of the chip results and comparison with other previous reported are shown in Table 1. This work achieves excellent prediction accuracy and energy-efficient after the algorithm hardware mapping.

![Figure 8. The training and prediction results of the model.](image1)

![Figure 9. Layout photograph and test board of the proposed chip.](image2)

| Implementation | This Work | Ref [16] | Ref [19] | Ref [20] |
|----------------|-----------|-----------|-----------|-----------|
| Algorithm      | ASIC      | ASIC      | Software  | Software  |
| Accuracy       | 93.2%     | 74%       | 60%       | 95%       |
| Power (mW)     | 12        | 65        | N/A       | N/A       |
| Energy/Sample (μJ) | 10.3 | 56        | N/A       | N/A       |

*Bidirectional Gated Recurrent Unit
6. Conclusion
In this work, an energy-efficient hardware accelerator was implemented to mapping spiking-LSTM neural network, containing 256 neurons and 64k synapses in 0.96 $\text{mm}^2$ area. LIF model is realized without using the multipliers saving a large mount of chip area and power. The chip can achieve an energy efficiency of 10.3 $\text{fJ}$ per sample and a predicting accuracy of about 93.2% in a forecasting model using the stock price of Google from Yahoo finance.

References
[1] Silver D, Huang A, Maddison C J, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M and Dieleman S, 2016 Mastering the game of Go with deep neural networks and tree search Nature 529(7587) 484-489
[2] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D and Riedmiller M, 2013 Playing atari with deep reinforcement learning arXiv preprint arXiv:1312.5602
[3] Wang J J, Hu S G, Zhan XT, Yu Q, Liu Z, Chen TP, Yin Y, Hosaka S and Liu Y, 2018 Handwritten-digit recognition by hybrid convolutional neural network based on HfO 2 memristive spiking-neuron Scientific Reports 8(1) 1-7
[4] Krizhevsky A, Sutskever I and Hinton GE, 2017 Imagenet classification with deep convolutional neural networks Communications of the ACM, 60(6) 84-90
[5] Maggiori E, Tarabalka Y, Charpiat G and Alliez P, 2016 Convolutional neural networks for large-scale remote-sensing image classification IEEE Transactions on Geoscience and Remote Sensing 55(2) 645-657
[6] Kong D, Hu S, Wu Y, Wang J, Xiong C, Yu Q, Shi Z, Liu Z, Chen T, Yin Y and Hosaka S, 2018 Realization of a power-efficient transmitter based on integrated artificial neural network IEEE Access 6 68773-68781
[7] Xiong W, Wu L, Alleva F, Droppo J, Huang X and Stolcke A, 2018, April The Microsoft 2017 conversational speech recognition system In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) IEEE pp 5934-5938
[8] Devlin J, Chang MW, Lee K and Toutanova K, 2018 Bert: Pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:181004805
[9] Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I, 2019 Language models are unsupervised multitask learners OpenAI Blog 1(8) 9
[10] Ghosh-Dastidar S and Adeli H, 2009 Spiking neural networks International Journal of Neural Systems 19(4) 295-308
[11] Wang JJ, Yu Q, Hu SG, Liu Y, Guo R, Chen TP, Yin Y and Liu Y, 2019 Winner-takes-all mechanism realized by memristive neural network Applied Physics Letters 115(24) 243701
[12] Tapiador-Morales R, Linares-Barranco A, Jimenez-Fernandez A and Jimenez-Moreno G, 2018 Neuromorphic lif row-by-row multiconvolution processor for fpga IEEE Transactions on Biomedical Circuits and Systems 13(1) 159-169
[13] Soleimani H and Drakakis EM, 2017 An efficient and reconfigurable synchronous neuron model IEEE Transactions on Circuits and Systems II: Express Briefs 65(1) 91-95
[14] Akopyan F, Sawada J, Cassidy A, Alvarez-Icaza R, Arthur J, Merolla P, Imam N, Nakamura Y, Datta P, Nam GJ and Taba B, 2015 Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 34(10) 1537-1557
[15] Davies M, Srinivasa N, Lin TH, Chinya G, Cao Y, Choday SH, Dimou G, Joshi P, Imam N, Jain S and Liao Y, 2018 Loihi: A neuromorphic manycore processor with on-chip learning IEEE Micro 38(1) 82-99
[16] Diehl PU, Zarrella G, Cassidy A, Pedroni BU and Neftci E, 2016, October Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware In 2016 IEEE International Conference on Rebooting Computing (ICRC) IEEE pp 1-8
[17] Balaji A, Das A, Wu Y, Huynh K, Dell’Anna FG, Indiveri G, Krichmar JL, Dutt ND, Schaafsma S and Catthoor F, 2019 Mapping spiking neural networks to neuromorphic hardware *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 28(1) 76-86

[18] Sun W, Park J, Jo S, Lee J and Shin H, 2019, January Implementation of multi-layer neural network system for neuromorphic hardware architecture 2019 *International Conference on Electronics, Information, and Communication (ICEIC)* IEEE pp 1-2

[19] Huynh HD, Dang LM and Duong D, 2017 December A new model for stock price movements prediction using deep neural network *Proceedings of the Eighth International Symposium on Information and Communication Technology* pp 57-62

[20] Ding G and Qin L, 2020 Study on the prediction of stock price based on the associated network model of LSTM *International Journal of Machine Learning and Cybernetics* 11(6) 1307-1317

[21] Thorpe S, Delorme A and Van Rullen R, 2001 Spike-based strategies for rapid processing *Neural Networks* 14(6-7) 715-725