Kamyar Givaki

Ph.D. in Computer Engineering - Computer Systems Architecture

(+98) 938-639-6473 | givakik@gmail.com | givakik | Kamyar Givaki

Computer Arithmetic, Stochastic and approximate Computing, Low-Power Design, Machine Learning

School of Computer Science

Institute for Research in Fundamental Sciences (IPM)

First-year Research plan

Dear Selection Committee

Enhancing the Efficiency of Processing ML Workloads Using Stochastic Computing

Machine learning (ML) workloads are data-intensive and computationally demanding applications that need efficient computation methods. New models are recently been popular to be used in many real-world applications. Attention-based Neural Networks (Transformers), Graph Neural Networks (GNNs), and Spiking Neural Networks (SNNs) are dominant trends in today's applications. Stochastic computing (SC) uses probabilistic bit-streams and simple logic gates to perform arithmetic operations. SC Properties makes it a good fit to be used in designing hardware accelerators for all of these three models. Below I will describe possible research opportunities for each of the three models.

SC can be used in transformers in different parts and layers. The attention mechanism, employed in transformers allows the network to selectively focus on the most informative parts of the input data. Finding the most informative parts of the data at the beginning of computations can help to decrease the required computation load by pruning parts with less impact on the final results. SC circuits have two key properties: they occupy low areas and consume low power. These properties can be used to design an efficient and highly paralleled module (technique) for detecting weak attention parts of a given data and transformer. This module can be used to identify the most informative parts of the input data, which can then be used by other computations performed using conventional or SC-based methods. So, My first research line is to design this detection module regarding other properties of transformers to enable efficient hardware implementation of this type of network in constrained devices.

SNNs are artificial neural networks that mimic biological neurons, by sending and receiving signals only when they reach a certain threshold. These signals, called spikes, are pulses of electrical activity that carry information. SNNs need efficient and reliable spike generation, which can be hard to do in hardware, because of variations in timing and voltage of the input signals. These variations, called skew, can affect the accuracy and consistency of the spike generation. SC can help implement SNNs in hardware, using analog circuits that can handle skew effects. SC can deal with variations in the input signals, without changing the output results. Developing SNN architectures that leverage the unique skew-tolerance property of SC is an additional research endeavor for the initial year of my postdoctoral period.

Body Sensor Networks (BSNs) are wireless networks of wearable or implantable devices that can monitor various physiological and environmental parameters of the human body. They can be used for applications such as health care, fitness, and sports. GNNs learn from graph data, where nodes are entities and edges mimic relationships between entities. GNNs apply functions to the aggregated information of nodes and their neighbors to form a new node context, which is costly in power and energy for edge devices. This is repeated for several layers until the final output is produced. The context update for each node can be performed can be performed locally on the device. However, this process requires high power consumption, which may exceed the computing capability of BSN devices. SC lowers the cost of this process by offering low-power hardware elements. This way, SC can help improve the efficiency and performance of graph neural networks on edge devices, especially in body implantable devices. Therefore, an architecture for updating the node's context efficiently is still a requirement and is an important and challenging problem for both computer and biology communities.

Main Objectives and expected outcomes of my research

The low power and low area demand of SC circuits makes them a promising solution to the limited power and area budget devices. My research aims to use SC to enhance the computation efficiency of the aforementioned ML workloads. My research plan consists of four main objectives:

1. A novel SC-based technique for detecting weak attentions will be proposed, which can reduce the computation load and power consumption of the network. The research will also explore novel SNN architectures that take advantage of the skew-tolerance property of Stochastic Computing, and evaluate their performance and accuracy on various tasks. The research will contribute to the advancement of neuromorphic computing, which aims to emulate the brain's functionality and efficiency.

2. Designing efficient and reliable spike generators for SNNs in hardware based on the skew-tolerance property of SC elements. The research will also evaluate the performance and efficiency of the proposed module in comparison with conventional or stochastic computing-based methods for different types of transformers and applications.

3. Novel architecture for updating the node's context in GNNs will be designed, which can reduce the power consumption of BSN devices. The research will also evaluate the accuracy and efficiency of the proposed architecture on various BSN applications, such as health care, fitness, and sports. The research will contribute to the advancement of edge computing and wireless sensor networks, which can enable new possibilities for human-machine interaction.

4. Publications in reputable conferences in the field of computer architecture.

5. Patents on the novel SC methods and architectures optimized for the aforementioned ML workloads.

Sincerely,

Kamyar Givaki