Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M18-121P
Published: 8/6/2019

Inventors

Elham Azari

Sarma Vrudhula

Contact

Shen Yan
shen.yan@skysonginnovations.com

An Energy Efficient Hardware Accelerator for an LSTM-RNN

Background

Energy-efficient portable devices, such as smartphones and wearable electronics, now dominate the personal computing market. As artificial intelligence continues to develop, integration of deep learning into embedded and mobile applications has become a natural course in its evolutionary. Unlike deep neural networks (DNNs) and convolutional neural networks (CNNs), recurrent neural networks (RNNs) feature feedback connections and are better suited for sequence learning. This makes RNNs desirable for processing the time-varying data sequences collected by many mobile systems. However, effective hardware implementation of RNNs remains inflexible due to the very large number of parameters and compute-intensive operations involved. As a result, RNN designs using conventional architectures require large area (i.e., number of cells) and high power consumption, which present challenges for embedded applications.

Invention Description

Researchers at Arizona State University have developed a low-power, energy-efficient ASIC designed for scalable RNNs in embedded devices. Specifically, this innovation leverages the Long Short-Term Memory (LSTM) model of RNN because of its strengths in handling classification tasks involving long-term temporal dependencies including handwriting recognition and speech-to-text translation. The compute-intensive units of LSTM—matrix-vector multiplication (MVM) and element-wise multiplier (EM)—are optimized to reduce area and power requirements. An accurate state-of-the-art stochastic computing (SC) multiplier is used within a multiply-and-accumulate (MAC) unit, providing resource-preserving MVM operation. To maximize throughput, this highly pipelined design features hierarchical controllers that synchronize variable-cycle multiplication operations with single-cycle units.

Implementation of the design on a Xilinx Zynq XC7Z030 FPGA for language modeling outperforms state-of-the-art designs by up to 27.9x, 7.7x, and 11.1x in throughput, and up to 45.3x, 14.8x, and 17.0x in energy efficiency.

Potential Applications

- Energy-efficient deep learning
- Data classification
- Language processing

Benefits and Advantages

- Efficient – Significantly reduces area utilization and power consumption in LSTM RNNs.
- High-Performance – Delivers up to 27.9x the throughput of competing language processing implementations.
- Versatile – Design architecture is applicable to all types of LSTM neural networks.

Homepage of Professor Sarma Vrudhula