IDEC Chip Design Contest

Energy-efficient configurable LSTM accelerator for mobile systems

Junseo Jo, Gunho Park, Jongmin Park and Youngjoo Lee
Department of Electrical Engineering
Pohang University of Science and Technology

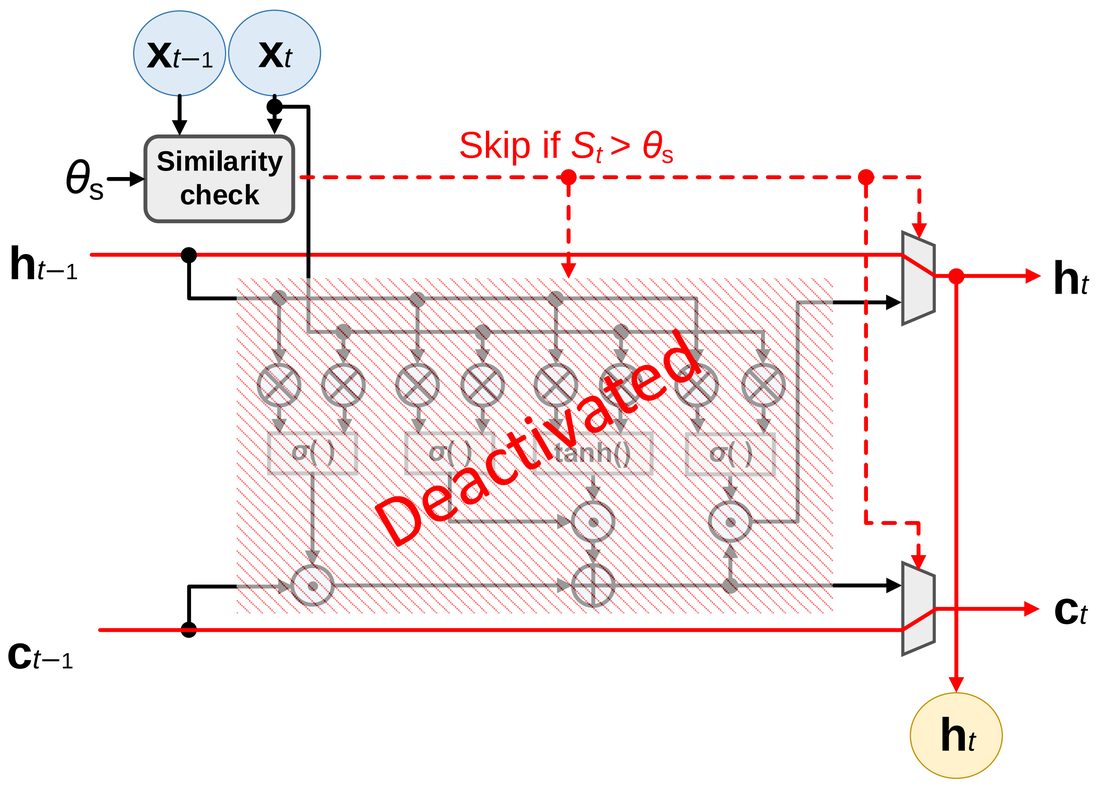
POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Abstract

This project proposes an accuracy configurable energy-efficient approximate recurrent neural network (RNN) based long short-term memory (LSTM) processor for mobile systems. The conventional RNN-based systems have been suffered from inefficient processing which caused by recurrent processing fashion. With the conventional approach, most of previous studies only focused on improving processing speed of system, which had been applied on servers and cloud systems. In recent studies, attempt on LSTM optimization design technique has been approved by increasing demand on AI-based system. For implementing effective throughput with energy efficient processing, the bit precision-based operation can provide reasonable energy efficiency with the proper accuracy for optimized mobile systems. As mentioned above, the bit precision-based method will also control accuracy in real-time processing. The proposed LSTM processor can achieve state-of-the-art energy efficient technique with the scaled accuracy. All the functionalities of the proposed work are verified by software simulations using Tensorflow and verified by RTL simulation using Xilinx FPGA. Also, we will synthesize the proposed hardware with power-aware P&R. As a result, the proposed system is fabricated in 65nm CMOS process, this system could be realization for state-of-the art NN based processor hardware design.

Motivation

❖ Similarity-based LSTM architecture



- Comparing two adjacent input data of LSTM network
- 3 modes (normal, delta, skip) according to similarity scores
  - Normal mode with high accuracy & energy usage
  - Delta mode with relatively low accuracy & energy usage
  - Skip mode skips all LSTM operation

Evaluation

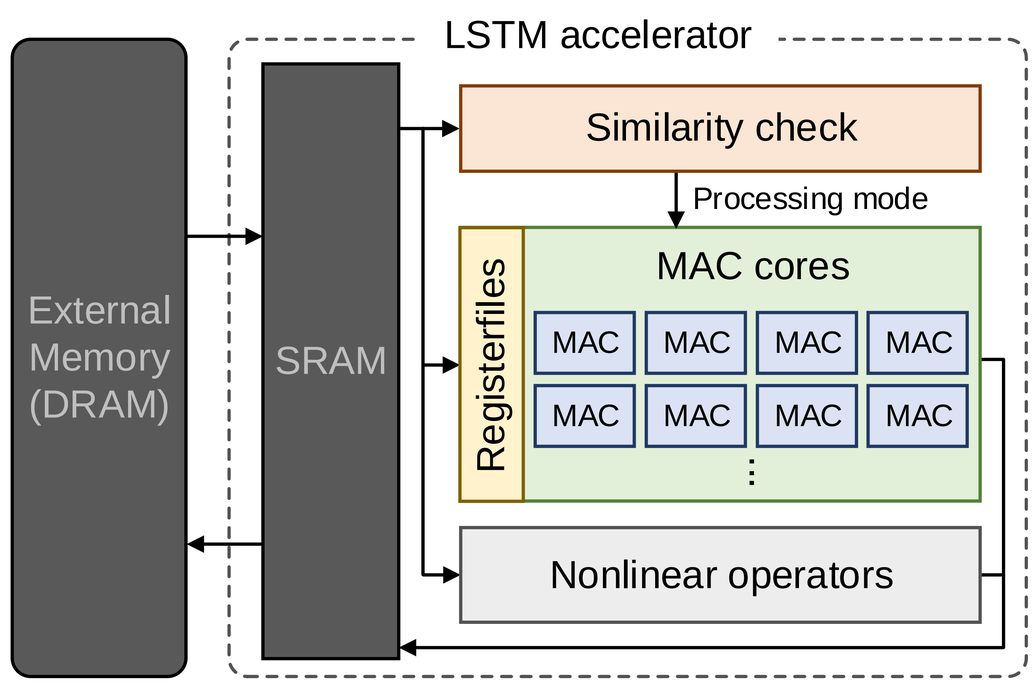
❖ Synthesis report

- Power consumption: 33.14 mW @1.2 V supply voltage
- Maximum throughput: 64 GOPS
- Energy efficiency: 1.93 TOPS / W
- Reducing 55% of total operations and 55% of DRAM access

This work was supported by the IC Design Education Center (IDEC) and the Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-TB1703-07.

Overall Architecture

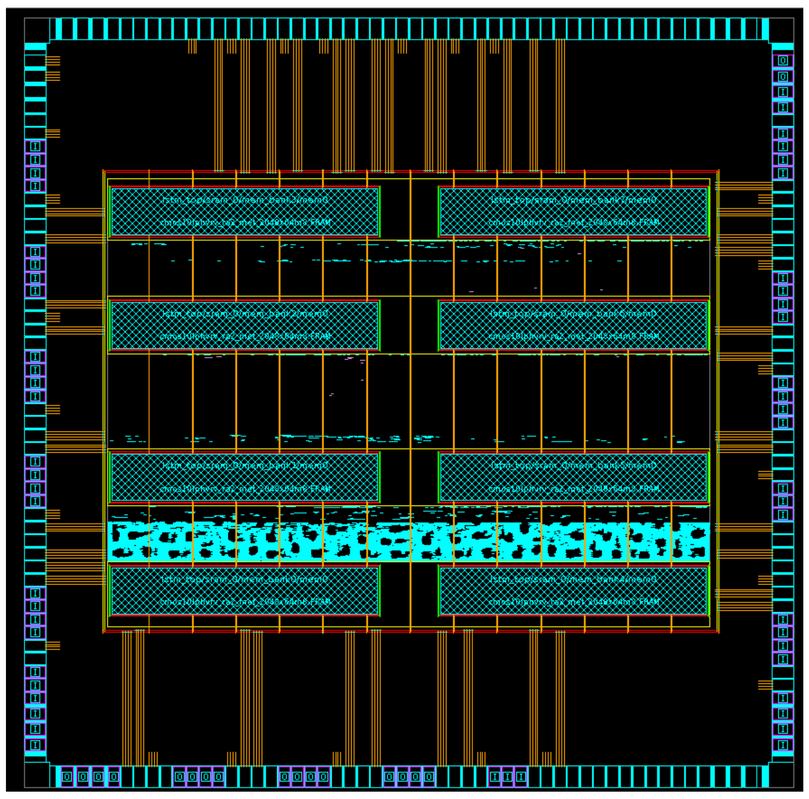
❖ Energy-efficient configurable LSTM accelerator



- Approximate multiplier for MAC cores

Implementation Results

❖ Die-photo of the accelerator



반도체설계교육센터
IC DESIGN EDUCATION CENTER