InnovateFPGA

PR144

SC-DNN
Deep Neural Network using Stochastic Computing

Po-Chun Chien, Yi-Hsuan Wu, Jung-Cheng Ho

Department of Electrical Engineering, National Taiwan University

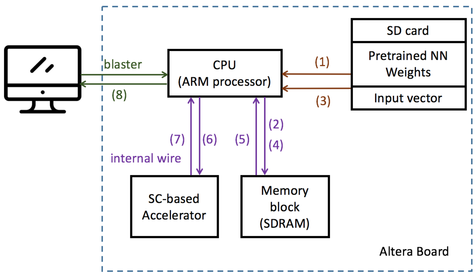
Introduction

The deep neural network is currently the most powerful machine learning technique. However, it requires high computation effort, especially in MAC operations, which is not applicable to smaller devices with power constraints.

Stochastic computation requires extremely low cost and power consumption compared to conventional fixed-point arithmetic. However, it comes with random errors and longer latency. In this project, our team is trying to find out a more powerful stochastic computation method to shorten the latency while not sacrificing the accuracy.

System Architecture

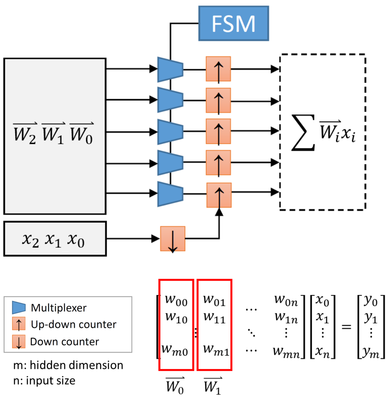
1. Store the pre-trained model parameters in the SD card.
2. CPU reads W and x into SDRAM.
3. CPU passes a column of W read from SDRAM and the current x into our IP.
4. The CPU receives the result calculated by the matrix-vector multiplier and repeats step 3 until the process is completed



SC Matric-Vector Multiplier

Stochastic Computation (SC) :
For any number between 0 and 1, we can convert it into a random binary bit stream, where the probability of 1 is its magnitude. The advantages of SC are (1)less gate (2)Low power (3)High error (bit-flip) tolerance.

Matric-Vector Multiplier
The structure was proposed by Hyeonuk Sim at DAC 2017. X is the input vector, will connect to a down counter, which will decide when multiplications will end. W is the weight matrix, a column will pass through MUXs of X to decide which bit stream to output.

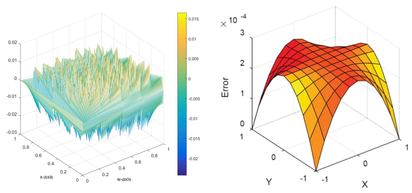


Results

Model Description: 784 → 50 → 50 → 50 → 10, activation = tanh
Each element in weight matrixes and input vector is limited to -1 to 1, with 8-bit precision.

Error Analysis

Error Surface



Our SC Conventional SC

Comparison

Platform	Accuracy
FPGA	91.07%
software	94.41%

A significant 3% drop in accuracy!!
Can probably be fixed if we use longer bit-streams for computation.

Latency Analysis

Theoretical value :
Without considering CPU operation time, the total time of data transmission by AXI bus and SC computation would be 2.63ms

Experimental results: the average computation time is 15.22ms, 65~66 images/sec.

Operation	load	Layer1		Layer2		Layer3		Layer4		other	total
		write	read	write	read	write	read	write	read		
Time (ms)	0.34	12.0	0.056	0.84	0.052	0.81	0.051	0.84	0.051	0.21	15.22
(%)	2.23	78.8	0.37	5.52	0.34	5.32	0.34	5.52	0.34	1.38	100