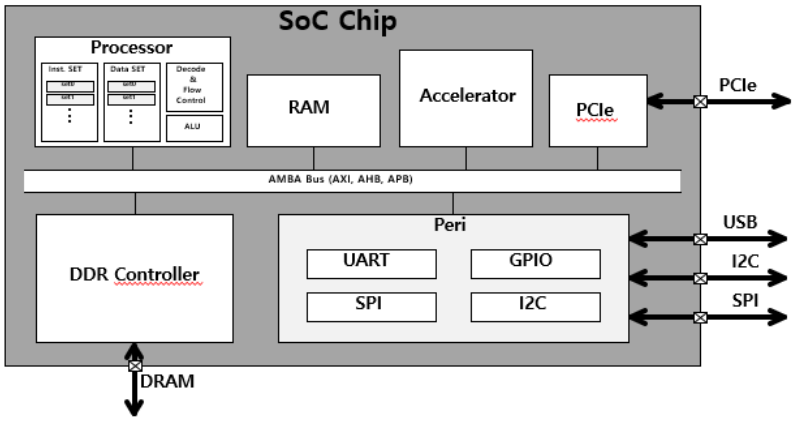
Lab 5. GEMM Accelerator

I. Purpose

The purpose of this lab is to design and implement a simple GEMM accelerator for matrix multiplication, accelerator the matrix multiplication on real FPGA hardware.

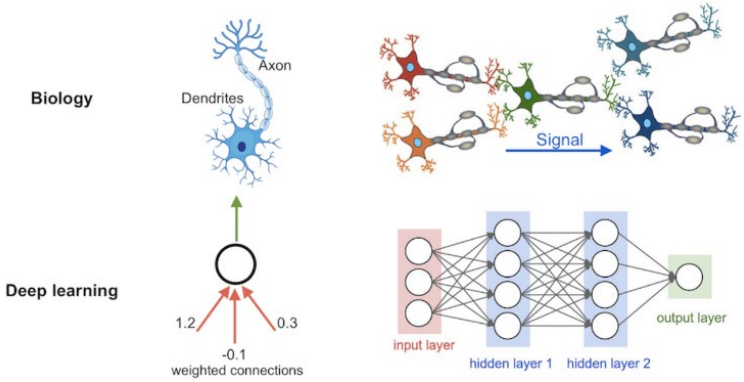
II. Backgrounds

A. System-on-Chip (SoC)



SoC is an integrated circuit that integrates all or most components of a computer or other electronic system. These components always include a CPU, memory, I/O ports. SoC controls its components by the bus interconnection, such as AXI and ABP protocols. The memory-mapped system is used for managing with address space which is commonly used in embedded system.

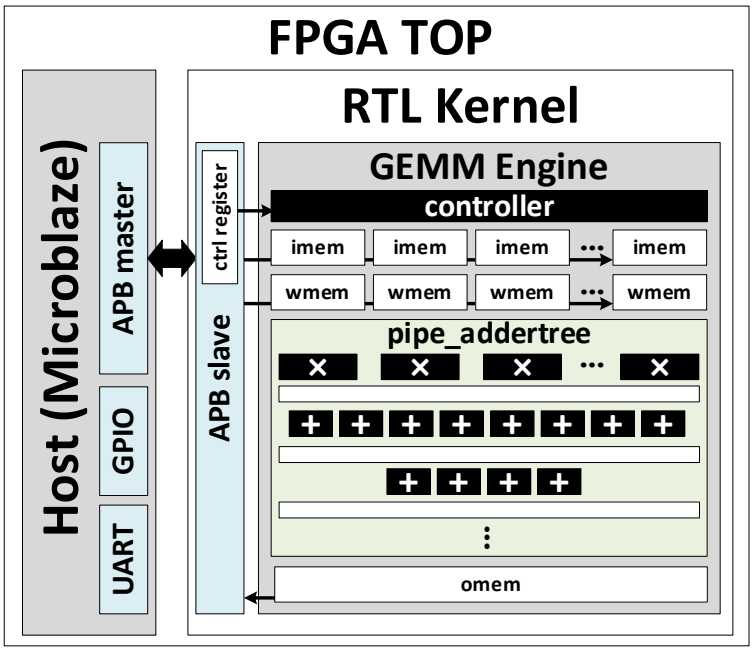
B. Deep Neural Network



Artificial neural networks are algorithms that imitate how the human brain recognizes patterns. We can use artificial neural networks to recognize specific patterns in diverse visual and auditory input data and use them in applications such as object detection and image classification. A deep neural network (DNN) is a neural network consisting of several hidden layers, as shown in Figure 1. Several nodes in each layer are designed to simulate the process occurring in the neurons. A node reacts when stimulated to a certain threshold, and this response magnitude is proportional to the input value and the coefficient of each node (referred to as weight). Generally, a node receives multiple inputs and has as many weights as the number of inputs. As a result of updating this weight, each input will be treated differently, and the output values of nodes can eventually be utilized for classification and regression analysis based on which input nodes they focus

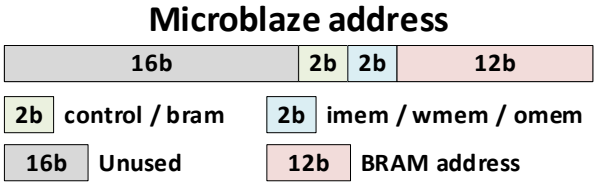
on. Depending on the computation type, the layers constituting the deep neural network can be classified in various ways, such as a convolution layer, a fully connected layer, and an activation layer. This deep neural network learns different characteristics of input layer by layer, and this process of abstraction allows us to understand and analyze even very large, high-dimensional data.

III. Lab Procedure



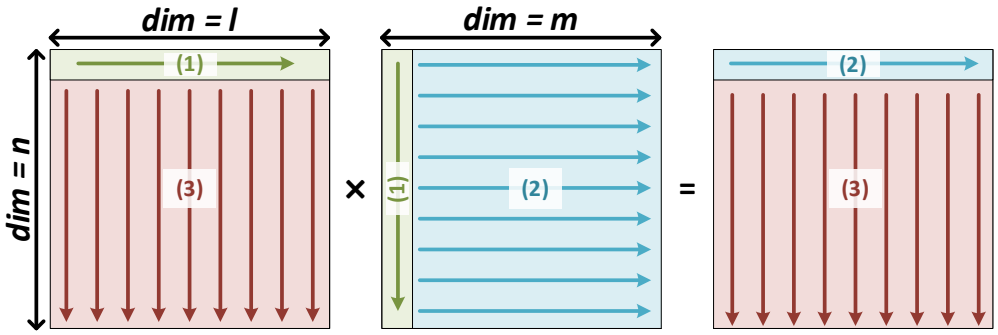
<Overall architecture of GEMM Engine SoC>

Problem 5A. Design a memory-mapped system



1) Design a memory-mapped system with given address decoding (16bit)
   - **A.** Select control register or memory access by green bit
   - **B.** Select which memory to access (imem, wmem, and omem) by blue bit
   - **C.** BRAM address bit within a single memory
2) Modify the apb_slave.sv file from Lab3 and implement the Microblaze-BRAM interface
3) Verify the demo of the memory access in your memory-mapped system

Problem 5B. Design GEMM accelerator's datapath & controller



1) Design GEMM accelerator's datapath (vector type)

2) Design the controller that controls the datapath (memory + ALU) for running GEMM through iterations

3) Verify GEMM accelerator's operation in simulation

Problem 5C. Verify & Demo GEMM accelerator on FPGA

1) Integrate all parts and verify the GEMM operation on FPGA

For design details, please refer to the provided supplemental material

IV. Final Report

Followings **should** be included in the report

1. Screen capture of simulation result of memory-mapped system
2. Screen capture of simulation result of GEMM accelerator
3. Screen capture of the synthesis and implementation of GEMM accelerator
4. Screen capture of register setting (host – RTL kernel communication process)
5. Discussion for the following questions

    1) Describe your memory-mapped system with address field.

    2) Describe each module of your own RTL code.

    3) Our GEMM accelerator proceeds the operation with a single adder-tree. Describe how order of the matrix multiplication changes when adding an extra adder-tree to increase parallelism of the accelerator.

    4) Due to the limit on the number of BRAMs in Arty A7, the size of the matrix that can be calculated at once is limited. Explain how the host code should be change when a larger matrix operation is performed using the current accelerator architecture. Also, explain the order in which input, weight, and output should be loaded to the GEMM accelerator.

V. References