MLPERF-TINY-CFU ACCELERATOR

Hua-Chen Wu

Agenda

Introduction

Hardware Design

Hardware-Software Co-Design

Evaluation

Introduction

The original design of the MLPerf™ Tiny image classification benchmark model shows **conv_2D** as the most time-consuming operation, with a total inference latency of 3,975,322.175 µs (**approximately 3.98 s**)

```
"Event","Tag","Ticks"
0,CONV_2D,17821
1,CONV_2D,54559
2,CONV_2D,51742
3,ADD,4010
4,CONV_2D,28776
5,CONV_2D,49949
6,CONV_2D,4453
7,ADD,2051
8,CONV_2D,25770
9,CONV_2D,46956
10,CONV_2D,3798
11,ADD,817
12,AVERAGE_POOL_2D,52
13,RESHAPE,3
14,FULLY_CONNECTED,14
15,SOFTMAX,18
Perf counters not enabled.
   298M (    297786777 )  cycles total
OK   Golden tests passed
```

```
100%|                                                  | 200/200 [18:27<00:00,  5.54s/it]
Accuracy: 0.875 %
Latency: 3975322.175 us
```

Hardware Design

A hardware accelerator has been implemented to perform efficient matrix multiplication. The entire design is composed of three main modules:

CFU

- Responsible for interacting with external interfaces, including command reception, data read/write operations, and executing matrix computation instructions.
- Manages the data flow and coordinates the internal TPU module to perform matrix calculations.

TPU

- A **16x16** tensor processing module with the core functionality of performing matrix multiplication.
- Includes a finite state machine (FSM) to handle the data loading (FEED) and computation (CALC) processes.
- Uses SRAM to emulate buffers for storing matrices A, B, and C.
- Implements matrix tiling operations, supporting the **M256K256N256 tiling** strategy, which decomposes large matrices into multiple 16x16 tiles for efficient computation.

Systolic Array

- The hardware core for matrix computation, utilizing a systolic array architecture.
- Implements a **16x16 network of Processing Elements** (PEs), where each PE performs multiplication and addition operations for individual matrix elements.

Hardware-Software Co-Design

A comprehensive hardware-software co-accelerated design, leveraging efficient data processing and hardware instructions to enable high-performance convolution calculations. The implementation includes:

Using **4-channel unrolling (Loop Unrolling)**, the **input tensor is transformed into a 2D matrix** with dynamically calculated dimensions and unified structure, converting convolution into matrix multiplication to reduce costs and improve efficiency.

Im2Col Optimization

Using **4-channel unrolling (Loop Unrolling)**, the **kernel is flattened into a 2D matri**x consistent with the Im2Col structure, reducing indexing complexity, enhancing processing performance, and simplifying subsequent matrix multiplication operations.

Kernel Flattening Optimization

Tiling technology is used to partition matrices for processing, reducing data volume and hardware pressure while improving cache hit rates and computational efficiency. The im2col, kernel, and resultmtx matrices are dynamically initialized to ensure the correctness and efficiency of the tiling operations.

Tiling and Hardware Acceleration

Custom CFU instructions are used to accelerate data transfer and matrix computation, reducing processor involvement and memory access frequency. Hardware-optimized matrix multiplication improves computational performance and enhances overall efficiency.

CFU Hardware Instruction Acceleration

Evaluation

The accelerated design of the MLPerf™ Tiny image classification benchmark model demonstrates significant time savings for the **conv_2D** operation. The total inference latency is reduced to **370,306.88 µs** (approximately **0.37 s**) compared to the original design's latency of **3,975,322.175 µs** (approximately **3.98 s**). This represents a speedup of approximately **10.74 times**, highlighting the efficiency of the optimized implementation.

```
"Event","Tag","Ticks"
0,CONV_2D,2338
1,CONV_2D,4280
2,CONV_2D,4507
3,ADD,4008
4,CONV_2D,1631
5,CONV_2D,2280
6,CONV_2D,893
7,ADD,2051
8,CONV_2D,1122
9,CONV_2D,2333
10,CONV_2D,490
11,ADD,818
12,AVERAGE_POOL_2D,48
13,RESHAPE,3
14,FULLY_CONNECTED,14
15,SOFTMAX,19
Perf counters not enabled.
    27M (     27496997 )  cycles total
OK   Golden tests passed
```

```
100%|                                                  | 200/200 [06:27<00:00,  1.94s/it]
Accuracy: 0.875 %
Latency: 370306.88 us
```

THANK YOU

Hua-Chen Wu

trista.cs11@nycu.edu.tw