MNIST Classification on FPGA

Arka Maity

MNIST Benchmark

- **Images of Handwritten Digits (28 x 28 pixes)**
- **60K Training Samples**
- **10K Test Samples**
- **Widely used for benchmarking Classifiers**

PYNQ Platform

- **SoC : Zynq 7020 series SoC**
  - **Processing System : Dual-core ARM Cortex-A9 @ 667 MHz**
  - **Programming Logic : 85K Programmable Logic Cells, 53.2K LUTs, 140 X 36KB BRAMs**
- **Memory: 512MB DDR3 / FLASH**
- **Design Tool : Vivado HLS, Vivado Synthesis and Place and Route Tools**

Classifier Performance (8K Batch size)

- **Classification Accuracy (Misclassification Rate %)**
- **Inference Latency (in ms)**

Classification Algorithms

- **Linear Classifiers**
  - **compute Ax + b during classification**
  - **find A and b during training**
    - **Weight Matrix**
    - **Offset Vector**
- **Non-Linear Classifiers (like Neural Network)**
- **We consider only Linear Classifiers here**

Overview of Classification in FPGA

```
stream in offset vector                  // T_offset (Can be statically stored)
stream in weight vector                  // T_weight (Can be statically stored)
for each sample in batch {               // T_tile
        stream in input                  //  T_input
        compute  the label (classify)    //  T_compute
        stream out output                //  T_output
}
```

Baseline Implementation

Linear Ridge Classification : alpha = 1.0

16x16 image size : 256 Features

Classification Accuracy (Python model) : 19.08%

8 bit inputs, 8 bit Classifier weights and 32 bit output (Labels and Offsets)

TILING : 1024

ARRAY PARTITION : 64

Baseline Implementation Performance

Classification Accuracy - 79.86 %

Inference Latency - 89.5ms

Speedup - 3.13x

Optimization I

Linear Ridge Classification : alpha = 140.0

12x12 image size : 144 Features

Classification Accuracy (Python model) : 18.48%

8 bit inputs, 8 bit Classifier weights and 32 bit output (Labels and Offsets)

TILING : 1024

ARRAY PARTITION : 72

Optimization I - Performance

Classification Accuracy - 80.53%

Inference Latency - 53.3ms

Speedup - 3.04x

Latency Breakdown

Component	Latency (Clock Cycles)	Iteration Latency	#(Iteration)
T_offset	22	-	
T_weight	3440	-	
T_tile	3260496	407562	8
T_input	352256	-	
T_compute	10245	-	
T_output	45056	-	

Optimization II

Linear Ridge Classification : alpha = 140.0

12x12 image size : 144 Features

Classification Accuracy (Python model) : 18.48%

8 bit inputs, 8 bit Classifier weights and 32 bit output (Labels and Offsets)

TILING : 1024

ARRAY PARTITION : 72

Pipelining Input, Output and Weight Streams

Latency Breakdown

Component	Latency (Clock Cycles)	Iteration Latency	#(Iteration)
T_offset	22	-	
T_weight	760	-	
T_tile	270472	33809	8
T_input	18432	-	
T_compute	10245	-	
T_output	5125	-	

Optimization II - Performance

Classification Accuracy - 80.53%

Inference Latency - 5.4ms

Speedup - 29.62x