CNN-IA 2022-06-28

Analyzer Report (Loop Blocking)

This report generated by Convolutional Neural Network Inference Analyzer (CNN-IA) to summarize the analysis needed to reach the optimal loop blocking for mlp_fc3_batch16 using restricted schedule space.

Memory Architecture:

| | L0 | L1 | L2 | L3 |
|----------------|--------|------|-------|-----------|
| Capacity | 4 | 16 | 65536 | 536870912 |
| Access cost | 0.0125 | 0.05 | 6.0 | 200.0 |
| Static cost | 0.0 | 0.0 | 0.0 | 0.0 |
| Parallel count | 1 | 256 | 1 | 1 |
| Parallel mode | 0 | 1 | 0 | 0 |
| Parallel cost | 0.0 | 2.0 | 0.0 | 0.0 |

Precision : 16

Minimum utilization : 0.0%

Outputs can be buffered by MAC : $\,0\,$

Replication to improve utilization: True

Glossary:

- Memory Levels: (L0, L1, L2, L3)
The smallest index the nearest to CPU.

- Loop Notations: (FX, FY, OX, OY, OC, IC, ON)

FX: FILTER WIDTH

FY: FILTER HEIGHT

OX : OUTPUT WIDTH

OY: OUTPUT HEIGHT

OC: OUTPUT CHANNEL

IC : INPUT CHANNEL

ON : BATCH

CNN-IA 2022-06-28

Map Configuration

Loop Blocking (factors):

| | L0 | L1 | L2 | L3 |
|----|-----|-----|-------|------|
| FX | 1.0 | 3.0 | 1.0 | 1.0 |
| FY | 1.0 | 1.0 | 1.0 | 1.0 |
| ох | 1.0 | 1.0 | 1.0 | 1.0 |
| OY | 1.0 | 1.0 | 1.0 | 1.0 |
| ос | 1.0 | 2.0 | 1.0 | 50.0 |
| IC | 1.0 | 1.0 | 250.0 | 2.0 |
| ON | 1.0 | 2.0 | 1.0 | 8.0 |

Loop Partitioning (units):

| | L0 | L1 | L2 | L3 |
|----|-----|------|-----|-----|
| FX | 1.0 | 1.0 | 1.0 | 1.0 |
| FY | 1.0 | 3.0 | 1.0 | 1.0 |
| ОХ | 1.0 | 1.0 | 1.0 | 1.0 |
| OY | 1.0 | 13.0 | 1.0 | 1.0 |
| ОС | 1.0 | 4.0 | 1.0 | 1.0 |
| IC | 1.0 | 1.0 | 1.0 | 1.0 |
| ON | 1.0 | 1.0 | 1.0 | 1.0 |

Loop Ordering (from the innermost):

| | L0 | L1 | L2 | L3 |
|----|-----|-----|-----|-----|
| FX | 6.0 | 0.0 | 6.0 | 6.0 |
| FY | 6.0 | 1.0 | 6.0 | 6.0 |
| ох | 6.0 | 6.0 | 6.0 | 6.0 |
| OY | 6.0 | 3.0 | 6.0 | 6.0 |
| ос | 6.0 | 2.0 | 6.0 | 0.0 |
| IC | 6.0 | 6.0 | 0.0 | 1.0 |
| ON | 6.0 | 4.0 | 6.0 | 2.0 |

(Hinted schedule configurations are in green)

CNN-IA 2022-06-28

Schedule

Cost

| MEM | ENERGY (PJ) |
|---------|-------------|
| LO | 99950.0 |
| L1 | 2139800.0 |
| L1-PARA | 35648000.0 |
| L2 | 8472000.0 |
| TOTAL | 46359750.0 |

- L3 memory was not checked for invalid underutilized.