1. **Modify the C code into OpenCL versions (buffers and w/o buffers).**

*Problems encountered:*

1. In the host program, forgot to add a float conversion when generating the input/weight matrix, so that all of the elements in the output matrix are 0s and the result verification always reports result correct.
2. For the version without buffers, forgot to initialize the output matrix to 0s. Although emulation passes correctly, hardware complained and produces incorrect result.
3. For the version without buffers, ‘restrict’ keyword has to be removed from input, weights and output in the function declaration of the CNN kernel due to a bug in version 18.1.
4. **Due to the existence of cache, start to go with the version without buffers.**

Tm=32 Tr=Tc=4 4481.5ms 215MHz 16+2 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 0% | 0.01 | 0.5 | 59.4 | 6.4 |
| Input | 0% | 0.01 | 0.5 | 59.4 | 6.44 |
| output | 0% | 0.01 | 0.1 | 0.5 | 20 |

Running\_sum

Tm=32 Tr=Tc=4 70.629ms 257MHz 16+1 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 0% | 9.14 | 23.2 | 3804.6 | 6.4 |
| Input | 0% | 5.39 | 23.2 | 3800.3 | 25.03 |
| output | - | 0.08 | 0.2 | 29.8 | 6.38 |

1. **Due to the nature that kernel with single work item is able to pipeline its loops, start to work on optimizing the CNN kernel with single work item.**

No restrict

Tm=32 Tr=Tc=4 793.783ms 273MHz 16+2 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 0% | 0 | 2 | 338.8 | 100 |
| Input | 0% | 0 | 2 | 402.3 | 21.05 |
| output | - | 0 | 0.1 | 2.7 | 7.41 |

Restrict

Tm=32 Tr=Tc=4 21.111ms 244MHz 16+2 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 0% | 13.82 | 82.4 | 12859.1 | 100 |
| Input | 100% | 0.2 | 82.4 | 8.2 | 100 |
| output | - | 0.79 | 0.7 | 100.5 | 6.27 |

Restrict

Tm=9 Tr=Tc=8 20.886ms 210MHz 32+4 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 76.9% | 0.02 | 48 | 1491.2 | 100 |
| Input | 99.6% | 0.01 | 48 | 56.5 | 100 |
| output | - | 0.01 | 0.4 | 50.4 | 6.35 |

1. **Added #pragma unroll N to the N\_ifm loop.**

#pragma unroll 4

Tm=9 Tr=Tc=8 5.491ms 218MHz 64+8 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 98.5% | 0.15 | 88.9 | 788.4 | 100 |
| Input | 100% | 0.07 | 88.9 | 7.7 | 100 |
| output | - | 0.01 | 2.8 | 387.4 | 6.27 |

#pragma unroll 8

Tm=9 Tr=Tc=8 4.347ms 196MHz 128+16 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 98.4% | 0.37 | 62.2 | 1002.5 | 100 |
| Input | 100% | 0.16 | 62.2 | 4.9 | 100 |
| output | - | 0.01 | 3.9 | 487 | 6.26 |

#pragma unroll 16

Tm=9 Tr=Tc=8 4.104ms 146MHz 256+32 DSPs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cache Hit% | Stall% | Occupancy% | Bandwidth (MB/s) | Bandwidth Efficiency% |
| Weights | 99.6% | 0.28 | 44.8 | 279.9 | 100 |
| Input | 100% | 0.12 | 44.8 | 2.5 | 100 |
| output | - | 0 | 5.6 | 522.9 | 6.25 |