**Special Topics on Artificial Intelligence**

Heejun, Jung

20204065

Jungheejun93@gm.gist.ac.kr

1. Experimental Environments

I conducted fully connected models on following environments

|  |  |
| --- | --- |
| Device | Specification |
| SW | Windows 10 Homle |
| Processor | Intel(R) Core i7-9700 CPU @ 3.00GHZ 8 cores |
| System | 64-bit |
| Cache Memory | Size: 12288 |
| DRAM | 2666MHz / 64GB |

Table 1. Devices specification

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W = 1024 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg |
| Batch size |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 4.000 | 4.145 | 4.411 | 4.203 | 5.031 | 4.587 | 4.528 | 4.170 | 4.187 | 4.153 | 4.342 |
| 2 | 3.120 | 3.130 | 3.308 | 3.217 | 3.300 | 3.243 | 3.235 | 3.509 | 3.325 | 3.207 | 3.259 |
| 8 | 1.660 | 1.621 | 1.632 | 1.614 | 1.649 | 1.620 | 1.651 | 1.606 | 1.631 | 1.596 | 1.628 |
| 16 | 1.262 | 1.274 | 1.272 | 1.371 | 1.305 | 1.275 | 1.255 | 1.298 | 1.245 | 1.272 | 1.283 |
| 32 | 1.067 | 1.074 | 1.054 | 1.076 | 1.105 | 1.069 | 1.115 | 1.116 | 1.074 | 1.089 | 1.084 |
| 64 | 1.010 | 1.045 | 1.027 | 1.007 | 1.015 | 1.016 | 1.020 | 1.024 | 1.025 | 1.006 | 1.019 |

1. Measure Inference Time

Table 2. Inference Time correspond to 1024 width fully connected model (unit is second)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| W = 256 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Avg |
| Batch size |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 1.890 | 1.908 | 1.838 | 1.890 | 1.890 | 1.895 | 1.893 | 1.879 | 1.883 | 1.881 | 1.885 |
| 2 | 1.551 | 1.547 | 1.450 | 1.451 | 1.549 | 1.475 | 1.452 | 1.577 | 1.416 | 1.459 | 1.493 |
| 8 | 0.936 | 0.972 | 0.939 | 0.942 | 0.923 | 0.999 | 0.931 | 0.970 | 0.946 | 0.977 | 0.953 |
| 16 | 0.862 | 0.844 | 0.904 | 0.847 | 0.887 | 0.897 | 0.877 | 0.882 | 0.850 | 0.861 | 0.871 |
| 32 | 0.795 | 0.779 | 0.811 | 0.806 | 0.793 | 0.814 | 0.797 | 0.805 | 0.801 | 0.797 | 0.800 |
| 64 | 0.738 | 0.729 | 0.741 | 0.744 | 0.749 | 0.761 | 0.750 | 0.765 | 0.754 | 0.753 | 0.749 |

Table 3. Inference Time correspond to 256 width fully connected model (unit is second)

1. Analysis of Results

FC(Fully Connected) with 1024 width model takes about 3 to 4 times longer than model with 256 width because the number of parameters which is consist of each model affect to the number of operation to inference predictions.

Besides, under batch size 32 inference processing time is shown these process were on “peak memory bandwidth” line where introduced in [1]. On the other hand, over batch size 32 tests were on “peak floating-point performance”[1] because time interval between batch size 32 and 64 seems like flat on the line.

One of the conspicuous result is the larger batch size takes the shorter inference time. Because inference operation works on the CPU which has multithread on each core. Each thread is connected with DRAM and each batch size are fetched on each thread at once if CPU capacity is enough. Therefore, when batch size increase and achieve on specific number as 32 in this environment, inference time is reduced.

1. Reference

[1] Williams, Samuel, Andrew Waterman, and David Patterson. "Roofline: an insightful visual performance model for multicore architectures." Communications of the ACM 52.4 (2009): 65-76.