|  |  |  |  |
| --- | --- | --- | --- |
| **Llama-2-7b + DDP + 8 HPU** | llm-on-ray | llama-bench-script | optimum-habana |
| epoch | 3 | 3 | 3 |
| max\_memory\_allocated (GB) | 87.17 | 94.44 | 87.17 |
| memory\_allocated (GB) | 34.54 | 13.13 | 34.54 |
| total\_flos | 704574292GF | 704574292GF | 704574292GF |
| total\_memory\_available (GB) | 94.62 | 94.62 | 94.62 |
| train\_loss | 0.9511 | 0.9174 | 0.9412 |
| train\_runtime | 05:18.8 | 04:46.8 | 05:04.4 |
| train\_samples\_per\_second | 147.717 | 141.069 | 151.795 |
| train\_steps\_per\_second | 1.158 | 2.209 | 1.189 |
| batch size | 8 | 8 | 8 |
| eval\_accuracy |  | 0.7751 | 0.7741 |
| eval\_loss |  | 0.855 | 0.8597 |
| perplexity |  | 2.3513 | 2.3626 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Llama-2-70b + DEEPSPEED + 8 HPU** | llm-on-ray | llama-bench-script | optimum-habana |
| epoch |  | 2 | 2 |
| max\_memory\_allocated (GB) |  | 94.58 | 94.58 |
| memory\_allocated (GB) |  | 26.6 | 26.6 |
| total\_flos |  | 1264185GF | 1264185GF |
| total\_memory\_available (GB) |  | 94.62 | 94.62 |
| train\_loss |  | 0.873 | 0.873 |
| train\_runtime |  | 40:05.4 | 40:05.4 |
| train\_samples\_per\_second |  | 2.639 | 2.639 |
| train\_steps\_per\_second |  | 0.033 | 0.033 |
| batch size |  | 10 | 10 |
| eval\_accuracy |  | 0.7861 | 0.7862 |
| eval\_loss |  | 0.7942 | 0.7943 |
| perplexity |  | 2.2127 | 2.2129 |

**Llama-2-7b + DDP + 8 HPU:**

**Llm-on-ray:**

\*\*\*\*\* train metrics \*\*\*\*\*

epoch = 3.0

max\_memory\_allocated (GB) = 87.17

memory\_allocated (GB) = 34.54

total\_flos = 704574292GF

total\_memory\_available (GB) = 94.62

train\_loss = 0.9511

train\_runtime = 0:05:18.83

train\_samples\_per\_second = 147.717

train\_steps\_per\_second = 1.158

**Ray llama-bench script:**

\*\*\*\*\* train metrics \*\*\*\*\*

epoch = 3.0

max\_memory\_allocated (GB) = 94.44

memory\_allocated (GB) = 13.13

total\_flos = 704574292GF

total\_memory\_available (GB) = 94.62

train\_loss = 0.9174

train\_runtime = 0:04:46.77

train\_samples\_per\_second = 141.069

train\_steps\_per\_second = 2.209

**Optimum-habana:**

\*\*\*\*\* train metrics \*\*\*\*\*

epoch = 3.0

max\_memory\_allocated (GB) = 87.17

memory\_allocated (GB) = 34.54

total\_flos = 704574292GF

total\_memory\_available (GB) = 94.62

train\_loss = 0.9412

train\_runtime = 0:05:04.43

train\_samples\_per\_second = 151.795

train\_steps\_per\_second = 1.189

**Llama-2-70b + DEEPSPEED + 8 HPU:**

**Ray llama-bench script:**

\*\*\*\*\* train metrics \*\*\*\*\*

epoch = 2.0

max\_memory\_allocated (GB) = 94.51

memory\_allocated (GB) = 26.6

total\_flos = 1264185GF

total\_memory\_available (GB) = 94.62

train\_loss = 0.8729

train\_runtime = 0:40:05.71

train\_samples\_per\_second = 2.639

train\_steps\_per\_second = 0.033

**Optimum-habana:**

\*\*\*\*\* train metrics \*\*\*\*\*

epoch = 2.0

max\_memory\_allocated (GB) = 94.58

memory\_allocated (GB) = 26.6

total\_flos = 1264185GF

total\_memory\_available (GB) = 94.62

train\_loss = 0.873

train\_runtime = 0:40:05.40

train\_samples\_per\_second = 2.639

train\_steps\_per\_second = 0.033