|  |
| --- |
| **Transformer OP显存分析** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Module** | **OP** | **IN shape** | **OUT shape** | **memory\_allocated** | | **max\_memory\_allocated** | | **memory\_reserved** |
|  |  |  |  | **理论分析** | **实际消耗** | **理论分析** | **实际消耗** | **实际消耗** |
| Transformer self attn |  |  |  | **1667.9 MiB, 1707928.5 KiB** | **1667.9 MiB, 1707928.5 KiB** | **1682.9 MiB, 1723290.0 KiB** | **1682.9 MiB, 1723290.0 KiB** | **1756.0 MiB** |
| Layernorm | 2\*4096\*320 | 2\*4096\*320 | **1672.9 MiB, 1713048.2 KiB**  （2\*4096\*320)\*2/1024=5120 KiB  1707928.5+5120=1713048.5 KiB | **1672.9 MiB, 1713048.5 KiB** | **1682.9 MiB, 1723292.5 KiB** | **1682.9 MiB, 1723292.5 KiB** |
| to\_query\_linear | 2\*4096\*320 | 2\*4096\*320 | **1677.9 MiB, 1718168.5 KiB** | **1677.9 MiB, 1718168.5 KiB** | **1687.9 MiB, 1728412.5 KiB** | **1687.9 MiB, 1728412.5 KiB** |
| to\_key\_linear | 2\*4096\*320 | 2\*4096\*320 | **1682.9 MiB, 1723288.5 KiB** | **1682.9 MiB, 1723288.5 KiB** | **1692.9 MiB, 1733532.5 KiB** | **1692.9 MiB, 1733532.5 KiB** |
| to\_value\_linear | 2\*4096\*320 | 2\*4096\*320 | **1687.9 MiB, 1728408.5 KiB** | **1687.9 MiB, 1728408.5 KiB** | **1692.9 MiB, 1733532.5 KiB** | **1692.9 MiB, 1733532.5 KiB** |
| query\_rearrange | 2\*4096\*320 | 16\*4096\*40 | **1687.9 MiB, 1728408.5 KiB** | **1687.9 MiB, 1728408.5 KiB** | **1702.9 MiB, 1743774.5 KiB** | **1702.9 MiB, 1743774.5 KiB** |
| key\_rearrange | 2\*4096\*320 | 16\*4096\*40 | **1687.9 MiB, 1728408.5 KiB** | **1687.9 MiB, 1728408.5 KiB** | **1702.9 MiB, 1743774.5 KiB** | **1702.9 MiB, 1743774.5 KiB** |
| value\_rearrange | 2\*4096\*320 | 16\*4096\*40 | **1692.9 MiB, 1733528.5 KiB** | **1692.9 MiB, 1733528.5 KiB** | **1702.9 MiB, 1743774.5 KiB** | **1702.9 MiB, 1743774.5 KiB** |
| torch.empty | / | 16\*4096\*4096 | **2204.9 MiB, 2257816.5 KiB**  (16\*4096\*4096)\*2/1024=524288.0 KiB  1733528.5+524288.0=2257816.5 KiB | **2204.9 MiB, 2257816.5 KiB** | **2214.9 MiB，2268062.5 KiB** | **3248.9 MiB, 3306398.5 KiB** | **3292.0 MiB** |
| torch.baddbmm | 16\*4096\*40  16\*40\*4096 | 16\*4096\*4096 | **2716.9 MiB, 2782104.5 KiB**  (16\*4096\*4096)\*2/1024=524288.0 KiB  2257816.5+524288.0= 2782104.5 KiB | **2716.9 MiB, 2782104.5 KiB** | **2726.9 MiB，2792350.5 KiB** | **3248.9 MiB, 3306398.5 KiB** |
| softmax | 16\*4096\*4096 | 16\*4096\*4096 | **3228.9 MiB, 3306392.5 KiB**  (16\*4096\*4096)\*2/1024=524288.0 KiB  2782104.5+524288.0=3306392.5 KiB | **3228.9 MiB, 3306392.5 KiB** | **3238.9 MiB，3316638.5 KiB** | **3248.9 MiB, 3306398.5 KiB** |