|  |
| --- |
| **Pytorch SD Pipe@FP16显存分析** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Params** | **IN shape** | **OUT shape** | **memory\_allocated** | | **max\_memory\_allocated** | | | **memory\_reserved** | **cuda context** | | **nvidia-smi** | |
|  |  |  |  | **理论分析** | **实际消耗** | **激活值最大的block** | **理论分析** | **实际消耗** | **实际消耗** | **3090-cuda11.8** | **V100-cuda11.7** | **3090-cuda11.8** | **V100-cuda11.7** |
| **Load weigths** | 1321.4 M | / | / | **2650.2 MiB, 2713856.0 KiB**  1321.4\*2=2642.8 MiB | **2650.2 MiB, 2713856.0 KiB** | **/** | **2650.2 MiB, 2713856.0 KiB** | **2650.2 MiB, 2713856.0 KiB** | **2664.0 MiB** | **620 MiB** | **1127 MiB** | **3284 MiB** | **3791 MiB** |
| **Text encoder** | 123.1 M | 2\*77 | 2\*77\*768 | **2650.5 MiB, 2714087.0 KiB**  2\*77\*768\*2/1024=231 KiB  2713856.0+231=2714087.0 KiB | **2650.5 MiB, 2714087.0 KiB** | **text transformer ：clip attn**  （12\*77\*64+  12\*77\*64+  12\*77\*77\*6+  1\*77\*768\*6）  \*2/1024=1757.7 KiB | **2651.9 MiB, 2715613.7 KiB**  2713856.0+1757.7=2715613.7 KiB  max\_memory\_allocated=  2715613.7 KiB> 2713856.0 KiB | **2652.4 MiB, 2716071.5 KiB** | **2666.0 MiB** | **3286 MiB** | **3793 MiB** |
| **Unet** | 845.6 M | 2\*4\*64\*64  2\*77\*768 | 2\*4\*64\*64 | **2651.0 MiB, 2714599.0 KiB**  2\*4\*64\*64\*8\*2/1024=512 KiB  2714087.0+512=2714599.5 KiB | **2651.0 MiB, 2714599.5 KiB** | **transformer：self attn**  （10240+  2\*4096\*320+  2\*4096\*320+  16\*4096\*40+  16\*4096\*40+  16\*4096\*40+  16\*4096\*4096+  16\*4096\*4096+  16\*4096\*4096）\*2/1024=1608704.0 KiB | **4221.5 MiB，4322791.0 KiB**  2714087.0+1608704.0=4322791.0 KiB  max\_memory\_allocated=  4322791.0 KiB>2715613.7  KiB | **4241.6 MiB, 4343374.0 KiB** | **4770.0 MiB** | **5390 MiB** | **5897 MiB** |
| **Vae** | 79.8 M  (30.3+49.5) | 1\*4\*64\*64 | 1\*3\*512\*512 | **2651.0 MiB, 2714599.0 KiB**  cuda->cpu，因此未发生变化 | **2651.0 MiB, 2714599.5 KiB** | **vae decoder： attn**  (1\*512\*64\*64+  1\*512\*64\*64+  1\*512\*4096\*19+  1\*4096\*4096)\*2/1024/1024  =118784.0 KiB | **4221.5 MiB，4322791.0 KiB**  2714599.5+118784.0=2833383.5 KiB  max\_memory\_allocated=  2833383.5 KiB<4322791.0 KiB | **4241.6 MiB, 4343374.0 KiB** | **4770.0 MiB** | **5390 MiB** | **5897 MiB** |
| **Safe checker** | 272.9 M | 1\*3\*512\*512 | 1\*3\*512\*512 | 推理Pipe时仅load weigths但未推理，因此不参与计算 | **/** | **/** | **/** | **/** | **/** | **5390 MiB** | **5897 MiB** |