|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Grad. Accumulation | Grad. Checkpoint | Mixed Precision | LoRA |
| Memory | parameter | - | - | ↓ | ↓ |
| activation | - | ↓ | ↓ | ↓ |
| gradient | - | - | ↓ | - |
| optimizer state | ↓ | - | - | - |
| Computation | | ↓ | ↑ | ↓ | ↓ |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GA | OFF | | | | ON | | | |
| MP | OFF | | ON | | OFF | | ON | |
| LoRA | OFF | ON | OFF | ON | OFF | ON | OFF | ON |
| Peak Mem | - | 3.13×107 | - | 4.34×107 | - | 3.13×107 | - | 4.34×107 |
| Runtime | - | 256.750 | - | 90.911 | - | 253.861 | - | 90.630 |

8 layers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| GA | OFF | | ON | |
| MP | OFF | ON | OFF | ON |
| LoRA | OFF | | | |
| Peak Mem | 3.77×107 | 2.02×107 | 3.77×107 | 2.02×107 |
| Runtime | 214.698 | 68.415 | 125.985 | 56.547 |