Batch Size and Memory Usage Calculation

Memory Usage Components

Model Params + Optimizer States + Gradients + Activations

Activations Memory ∝ **Batch Size × Sequence Length**

Example: BERT-base

batch = 32 • **seq = 512** → **~8GB** activation memory

Gradient Accumulation

Split batch into micro-batches to save memory

Gradient Checkpointing

Recompute activations (trade compute for memory)

Rule of Thumb

OOM Error?

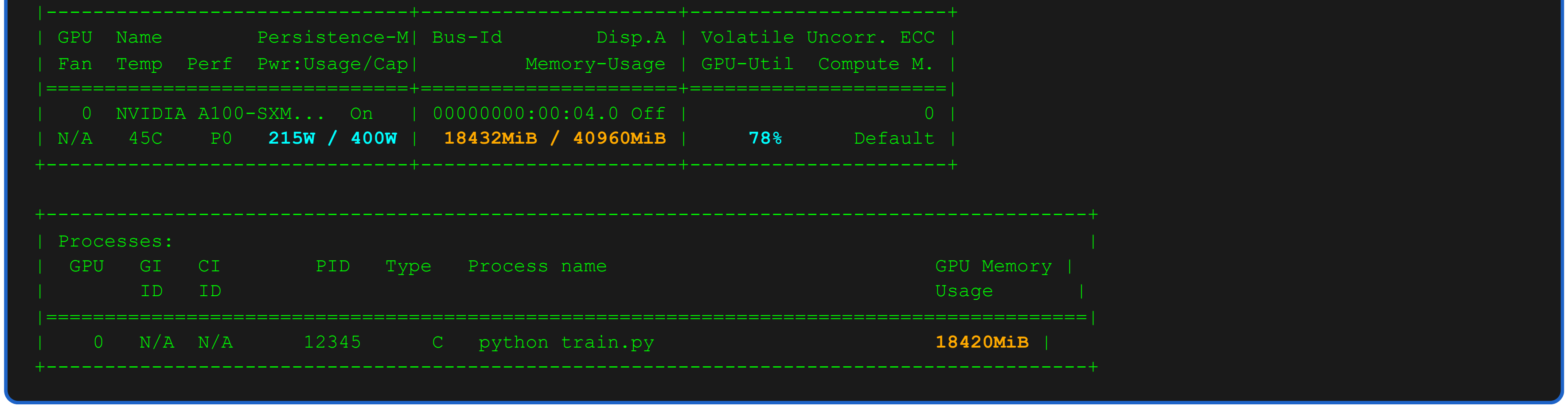
Reduce batch size by 50%

Monitoring Tools

`nvidia-smi`

`torch.cuda.memory_allocated()`

```
$ nvidia-smi
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.104.05   Driver Version: 535.104.05   CUDA Version: 12.2     |
```



```
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  NVIDIA A100-SXM...  On   | 00000000:00:04.0 Off |                    0 |
| N/A   45C    P0   215W / 400W |  18432MiB / 40960MiB |     78%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage       |
|=========================================================================================|
|    0   N/A  N/A      12345      C   python train.py                        18420MiB |
+-----------------------------------------------------------------------------------------+
```