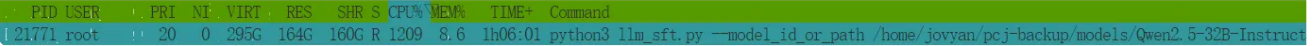


# 32b&72b模型Lora微调

## 32b-lora微调

1. 单节点单卡，配置deepspeed模式为零3-offload、使用flash\_attn、设置最大长度max\_length是16384，限制到 48G 显存( $48935\text{M}/1024 = 47.79\text{G}$ )，能够支持训练

2. 

3. 再进一步限制，限制到24G跑不起来，36G也可以勉强跑起来

4. 脚本

```
1  # 限制GPU使用量
2  import os
3  import torch
4  import torch.distributed as dist
5  import custom
6
7  from swift.llm import sft_main
8
9  if __name__ == '__main__':
10     # 配置单机单卡以及相关参数
11     os.environ['MASTER_ADDR'] = 'localhost'
12     os.environ['MASTER_PORT'] = '12355'
13     os.environ['CUDA_VISIBLE_DEVICES'] = '0'
14     os.environ['LOCAL_RANK'] = '0'
15     limits = 64    # GPU显存使用限制 单位GB
16
17     # 获取单颗GPU显存数量
18     total_memory = torch.cuda.get_device_properties(0).total_memory    # B
19     total_mem = total_memory / 1024 / 1024 / 1024    # GB
20
21     # 限制GPU使用量
22     ratio = limits / total_mem
23     torch.cuda.set_per_process_memory_fraction(ratio, 0)
24     dist.init_process_group("gloo", rank=0, world_size=1)
25     output = sft_main()
```

```
1  #!/bin/bash
2
3  nproc_per_node=1
4  CUDA_VISIBLE_DEVICES=0 \
5  NPROC_PER_NODE=$nproc_per_node \
6  python3 llm_sft_limit.py \
7      --model_id_or_path {32b基座模型地址}\
8      --model_type qwen2_5-32b-instruct \
9      --sft_type lora \
10     --tuner_backend peft \
11     --template_type AUTO \
12     --dtype AUTO \
13     --output_dir {输出目录} \
14     --ddp_backend nccl \
15     --dataset {训练集} \
16     --train_dataset_sample -1 \
17     --num_train_epochs 6 \
18     --max_length 16384 \
19     --check_dataset_strategy warning \
20     --lora_rank 64 \
21     --lora_alpha 16 \
22     --lora_dropout_p 0.05 \
23     --lora_target_modules ALL \
24     --gradient_checkpointing true \
25     --batch_size 1 \
26     --weight_decay 0.1 \
27     --learning_rate 1e-4 \
28     --gradient_accumulation_steps 2 \
29     --max_grad_norm 0.5 \
30     --warmup_ratio 0.03 \
31     --save_strategy epoch \
32     --evaluation_strategy epoch \
33     --logging_steps 10 \
34     --use_flash_attn true \
35     --save_total_limit 6 \
36     --deepspeed "zero3-offload"
```

## 72b-lora微调

1. 单节点单卡，配置deepspeed模式为零三-offload、使用flash\_attn、设置最大长度max\_length是16384，最低显存要求 64G，能够支持训练

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
2488	root	20	0	236G	120G	116G	R	100.	6.3	48h19:24	/home/jovyan/pcj-backup/anaconda3/envs/pcj/bin/python -u /home/jovyan/pcj-backup/workde
3100	root	20	0	171G	120G	116G	S	100.	6.3	50h34:26	/home/jovyan/pcj-backup/anaconda3/envs/pcj/bin/python -u /home/jovyan/pcj-backup/workde
10602	root	20	0	382G	231G	229G	R	100.	12.1	1h52:41	python3 llm_sft_limit.py --model_id_or_path /home/jovyan/pcj-backup/models/Qwen2.5-72B-

- 2.
3. 再进一步限制，48G是跑不起来的
4. 脚本

Plain Text

```

1  # 限制GPU使用量
2  import os
3  import torch
4  import torch.distributed as dist
5  import custom
6
7  from swift.llm import sft_main
8
9  if __name__ == '__main__':
10     # 配置单机单卡以及相关参数
11     os.environ['MASTER_ADDR'] = 'localhost'
12     os.environ['MASTER_PORT'] = '12355'
13     os.environ['CUDA_VISIBLE_DEVICES'] = '0'
14     os.environ['LOCAL_RANK'] = '0'
15     limits = 64    # GPU显存使用限制 单位GB
16
17     # 获取单颗GPU显存数量
18     total_memory = torch.cuda.get_device_properties(0).total_memory    # B
19     yte
20     total_mem = total_memory / 1024 / 1024 / 1024    # GB
21
22     # 限制GPU使用量
23     ratio = limits / total_mem
24     torch.cuda.set_per_process_memory_fraction(ratio, 0)
25     dist.init_process_group("gloo", rank=0, world_size=1)
26     output = sft_main()

```

```
1  #!/bin/bash
2
3  nproc_per_node=1
4  CUDA_VISIBLE_DEVICES=0 \
5  NPROC_PER_NODE=$nproc_per_node \
6  python3 llm_sft_limit.py \
7      --model_id_or_path {72b基座模型地址}\
8      --model_type qwen2_5-72b-instruct \
9      --sft_type lora \
10     --tuner_backend peft \
11     --template_type AUTO \
12     --dtype AUTO \
13     --output_dir {输出目录} \
14     --ddp_backend nccl \
15     --dataset {训练集} \
16     --train_dataset_sample -1 \
17     --num_train_epochs 6 \
18     --max_length 16384 \
19     --check_dataset_strategy warning \
20     --lora_rank 64 \
21     --lora_alpha 16 \
22     --lora_dropout_p 0.05 \
23     --lora_target_modules ALL \
24     --gradient_checkpointing true \
25     --batch_size 1 \
26     --weight_decay 0.1 \
27     --learning_rate 1e-4 \
28     --gradient_accumulation_steps 2 \
29     --max_grad_norm 0.5 \
30     --warmup_ratio 0.03 \
31     --save_strategy epoch \
32     --evaluation_strategy epoch \
33     --logging_steps 10 \
34     --use_flash_attn true \
35     --save_total_limit 6 \
36     --deepspeed "zero3-offload"
```