In [1]:

```
# pip install pytorch_memlab

# 其他需要的包
# pip install transformers thop torch

from transformers import AutoModelForCausalLM, AutoTokenizer
```

In [4]:

```
# 模型的snapshot本地地址
model_directory = r"Z:/llmfile/Llama-2-7B-hf/models--meta-llama--Llama-2-7b-hf/s
```

In [2]:

```
# 加载模型，这个用来确定能不能正确加载
# 成功后建议shut down kernal，重新import，直接运行下一个部分，否则会爆内存
model = AutoModelForCausalLM.from_pretrained(model_directory)
tokenizer = AutoTokenizer.from_pretrained(model_directory)

# 确保正确加载
print(model)
```

```
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(32000, 4096, padding_idx=0)
    (layers): ModuleList(
      (0-31): 32 x LlamaDecoderLayer(
        (self_attn): LlamaAttention(
          (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (k_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (v_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear(in_features=4096, out_features=11008, bias=False)
          (up_proj): Linear(in_features=4096, out_features=11008, bias=False)
          (down_proj): Linear(in_features=11008, out_features=4096, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): LlamaRMSNorm()
        (post_attention_layernorm): LlamaRMSNorm()
      )
    )
    (norm): LlamaRMSNorm()
  )
  (lm_head): Linear(in_features=4096, out_features=32000, bias=False)
)
```

In [5]:

```
from pytorch_memlab import MemReporter
```

In [6]:

```
model = AutoModelForCausalLM.from_pretrained(model_directory)
reporter = MemReporter(model)
reporter.report()  # 输出每层内存使用情况
```

```
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
```

```
Element type                                            Size  Used MEM
-------------------------------------------------------------------------------
Storage on cpu
Tensor0                                          (4096, 128)     2.00M
Tensor1                                          (4096, 128)     2.00M
Tensor2                                          (4096, 128)     2.00M
Tensor3                                          (4096, 128)     2.00M
Tensor4                                          (4096, 128)     2.00M
Tensor5                                          (4096, 128)     2.00M
Tensor6                                          (4096, 128)     2.00M
Tensor7                                          (4096, 128)     2.00M
Tensor8                                          (4096, 128)     2.00M
Tensor9                                          (4096, 128)     2.00M
Tensor10                                         (4096, 128)     2.00M
Tensor11                                         (4096, 128)     2.00M
Tensor12                                         (4096, 128)     2.00M
Tensor13                                         (4096, 128)     2.00M
Tensor14                                         (4096, 128)     2.00M
Tensor15                                         (4096, 128)     2.00M
Tensor16                                         (4096, 128)     2.00M
Tensor17                                         (4096, 128)     2.00M
Tensor18                                         (4096, 128)     2.00M
Tensor19                                         (4096, 128)     2.00M
Tensor20                                         (4096, 128)     2.00M
Tensor21                                         (4096, 128)     2.00M
Tensor22                                         (4096, 128)     2.00M
Tensor23                                         (4096, 128)     2.00M
Tensor24                                         (4096, 128)     2.00M
Tensor25                                         (4096, 128)     2.00M
Tensor26                                         (4096, 128)     2.00M
Tensor27                                         (4096, 128)     2.00M
Tensor28                                         (4096, 128)     2.00M
Tensor29                                         (4096, 128)     2.00M
Tensor30                                         (4096, 128)     2.00M
Tensor31                                         (4096, 128)     2.00M
Tensor32                                         (4096, 128)     2.00M
Tensor33                                         (4096, 128)     2.00M
Tensor34                                         (4096, 128)     2.00M
Tensor35                                         (4096, 128)     2.00M
Tensor36                                         (4096, 128)     2.00M
Tensor37                                         (4096, 128)     2.00M
Tensor38                                         (4096, 128)     2.00M
Tensor39                                         (4096, 128)     2.00M
Tensor40                                         (4096, 128)     2.00M
Tensor41                                         (4096, 128)     2.00M
Tensor42                                         (4096, 128)     2.00M
Tensor43                                         (4096, 128)     2.00M
Tensor44                                         (4096, 128)     2.00M
Tensor45                                         (4096, 128)     2.00M
Tensor46                                         (4096, 128)     2.00M
Tensor47                                         (4096, 128)     2.00M
Tensor48                                         (4096, 128)     2.00M
Tensor49                                         (4096, 128)     2.00M
Tensor50                                         (4096, 128)     2.00M
Tensor51                                         (4096, 128)     2.00M
Tensor52                                         (4096, 128)     2.00M
Tensor53                                         (4096, 128)     2.00M
Tensor54                                         (4096, 128)     2.00M
Tensor55                                         (4096, 128)     2.00M
Tensor56                                         (4096, 128)     2.00M
```

```
Tensor57                                         (4096, 128)     2.00M
Tensor58                                         (4096, 128)     2.00M
Tensor59                                         (4096, 128)     2.00M
Tensor60                                         (4096, 128)     2.00M
Tensor61                                         (4096, 128)     2.00M
Tensor62                                         (4096, 128)     2.00M
Tensor63                                         (4096, 128)     2.00M
lm_head.weight                                 (32000, 4096)   500.00M
model.embed_tokens.weight                      (32000, 4096)   500.00M
model.norm.weight                                    (4096,)    16.00K
model.layers.0.input_layernorm.weight                (4096,)    16.00K
model.layers.0.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.1.input_layernorm.weight                (4096,)    16.00K
model.layers.1.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.2.input_layernorm.weight                (4096,)    16.00K
model.layers.2.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.3.input_layernorm.weight                (4096,)    16.00K
model.layers.3.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.4.input_layernorm.weight                (4096,)    16.00K
model.layers.4.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.5.input_layernorm.weight                (4096,)    16.00K
model.layers.5.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.6.input_layernorm.weight                (4096,)    16.00K
model.layers.6.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.7.input_layernorm.weight                (4096,)    16.00K
model.layers.7.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.8.input_layernorm.weight                (4096,)    16.00K
model.layers.8.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.9.input_layernorm.weight                (4096,)    16.00K
model.layers.9.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.10.input_layernorm.weight               (4096,)    16.00K
model.layers.10.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.11.input_layernorm.weight               (4096,)    16.00K
model.layers.11.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.12.input_layernorm.weight               (4096,)    16.00K
model.layers.12.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.13.input_layernorm.weight               (4096,)    16.00K
model.layers.13.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.14.input_layernorm.weight               (4096,)    16.00K
model.layers.14.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.15.input_layernorm.weight               (4096,)    16.00K
model.layers.15.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.16.input_layernorm.weight               (4096,)    16.00K
model.layers.16.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.17.input_layernorm.weight               (4096,)    16.00K
model.layers.17.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.18.input_layernorm.weight               (4096,)    16.00K
model.layers.18.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.19.input_layernorm.weight               (4096,)    16.00K
model.layers.19.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.20.input_layernorm.weight               (4096,)    16.00K
model.layers.20.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.21.input_layernorm.weight               (4096,)    16.00K
model.layers.21.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.22.input_layernorm.weight               (4096,)    16.00K
model.layers.22.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.23.input_layernorm.weight               (4096,)    16.00K
model.layers.23.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.24.input_layernorm.weight               (4096,)    16.00K
model.layers.24.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.25.input_layernorm.weight               (4096,)    16.00K
model.layers.25.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.26.input_layernorm.weight               (4096,)    16.00K
model.layers.26.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.27.input_layernorm.weight               (4096,)    16.00K
model.layers.27.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.28.input_layernorm.weight               (4096,)    16.00K
model.layers.28.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.29.input_layernorm.weight               (4096,)    16.00K
model.layers.29.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.30.input_layernorm.weight               (4096,)    16.00K
model.layers.30.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.31.input_layernorm.weight               (4096,)    16.00K
model.layers.31.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.0.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.0.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.0.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.0.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor64                                               (64,)   512.00B
Tensor65                                   (1, 1, 4096, 128)     0.00B
Tensor66                                   (1, 1, 4096, 128)     0.00B
model.layers.0.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.0.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.0.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.1.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.1.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.1.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.1.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor67                                               (64,)   512.00B
Tensor68                                   (1, 1, 4096, 128)     0.00B
Tensor69                                   (1, 1, 4096, 128)     0.00B
model.layers.1.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.1.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.1.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.2.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.2.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.2.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.2.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor70                                               (64,)   512.00B
Tensor71                                   (1, 1, 4096, 128)     0.00B
Tensor72                                   (1, 1, 4096, 128)     0.00B
model.layers.2.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.2.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.2.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.3.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.3.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.3.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.3.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor73                                               (64,)   512.00B
Tensor74                                   (1, 1, 4096, 128)     0.00B
Tensor75                                   (1, 1, 4096, 128)     0.00B
model.layers.3.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.3.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.3.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.4.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.4.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.4.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.4.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor76                                               (64,)   512.00B
Tensor77                                   (1, 1, 4096, 128)     0.00B
```

```
Tensor78                                   (1, 1, 4096, 128)     0.00B
model.layers.4.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.4.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.4.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.5.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.5.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.5.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.5.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor79                                               (64,)   512.00B
Tensor80                                   (1, 1, 4096, 128)     0.00B
Tensor81                                   (1, 1, 4096, 128)     0.00B
model.layers.5.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.5.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.5.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.6.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.6.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.6.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.6.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor82                                               (64,)   512.00B
Tensor83                                   (1, 1, 4096, 128)     0.00B
Tensor84                                   (1, 1, 4096, 128)     0.00B
model.layers.6.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.6.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.6.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.7.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.7.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.7.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.7.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor85                                               (64,)   512.00B
Tensor86                                   (1, 1, 4096, 128)     0.00B
Tensor87                                   (1, 1, 4096, 128)     0.00B
model.layers.7.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.7.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.7.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.8.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.8.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.8.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.8.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor88                                               (64,)   512.00B
Tensor89                                   (1, 1, 4096, 128)     0.00B
Tensor90                                   (1, 1, 4096, 128)     0.00B
model.layers.8.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.8.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.8.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.9.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.9.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.9.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.9.self_attn.o_proj.weight          (4096, 4096)    64.00M
Tensor91                                               (64,)   512.00B
Tensor92                                   (1, 1, 4096, 128)     0.00B
Tensor93                                   (1, 1, 4096, 128)     0.00B
model.layers.9.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.9.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.9.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.10.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.10.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.10.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.10.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor94                                               (64,)   512.00B
Tensor95                                   (1, 1, 4096, 128)     0.00B
```

```
Tensor96                                   (1, 1, 4096, 128)     0.00B
model.layers.10.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.10.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.10.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.11.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.11.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.11.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.11.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor97                                               (64,)   512.00B
Tensor98                                   (1, 1, 4096, 128)     0.00B
Tensor99                                   (1, 1, 4096, 128)     0.00B
model.layers.11.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.11.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.11.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.12.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.12.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.12.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.12.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor100                                              (64,)   512.00B
Tensor101                                  (1, 1, 4096, 128)     0.00B
Tensor102                                  (1, 1, 4096, 128)     0.00B
model.layers.12.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.12.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.12.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.13.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.13.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.13.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.13.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor103                                              (64,)   512.00B
Tensor104                                  (1, 1, 4096, 128)     0.00B
Tensor105                                  (1, 1, 4096, 128)     0.00B
model.layers.13.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.13.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.13.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.14.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.14.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.14.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.14.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor106                                              (64,)   512.00B
Tensor107                                  (1, 1, 4096, 128)     0.00B
Tensor108                                  (1, 1, 4096, 128)     0.00B
model.layers.14.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.14.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.14.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.15.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.15.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.15.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.15.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor109                                              (64,)   512.00B
Tensor110                                  (1, 1, 4096, 128)     0.00B
Tensor111                                  (1, 1, 4096, 128)     0.00B
model.layers.15.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.15.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.15.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.16.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.16.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.16.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.16.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor112                                              (64,)   512.00B
Tensor113                                  (1, 1, 4096, 128)     0.00B
```

```
Tensor114                                  (1, 1, 4096, 128)     0.00B
model.layers.16.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.16.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.16.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.17.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.17.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.17.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.17.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor115                                              (64,)   512.00B
Tensor116                                  (1, 1, 4096, 128)     0.00B
Tensor117                                  (1, 1, 4096, 128)     0.00B
model.layers.17.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.17.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.17.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.18.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.18.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.18.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.18.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor118                                              (64,)   512.00B
Tensor119                                  (1, 1, 4096, 128)     0.00B
Tensor120                                  (1, 1, 4096, 128)     0.00B
model.layers.18.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.18.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.18.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.19.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.19.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.19.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.19.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor121                                              (64,)   512.00B
Tensor122                                  (1, 1, 4096, 128)     0.00B
Tensor123                                  (1, 1, 4096, 128)     0.00B
model.layers.19.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.19.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.19.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.20.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.20.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.20.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.20.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor124                                              (64,)   512.00B
Tensor125                                  (1, 1, 4096, 128)     0.00B
Tensor126                                  (1, 1, 4096, 128)     0.00B
model.layers.20.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.20.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.20.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.21.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.21.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.21.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.21.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor127                                              (64,)   512.00B
Tensor128                                  (1, 1, 4096, 128)     0.00B
Tensor129                                  (1, 1, 4096, 128)     0.00B
model.layers.21.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.21.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.21.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.22.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.22.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.22.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.22.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor130                                              (64,)   512.00B
Tensor131                                  (1, 1, 4096, 128)     0.00B
```

```
Tensor132                                  (1, 1, 4096, 128)     0.00B
model.layers.22.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.22.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.22.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.23.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.23.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.23.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.23.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor133                                              (64,)   512.00B
Tensor134                                  (1, 1, 4096, 128)     0.00B
Tensor135                                  (1, 1, 4096, 128)     0.00B
model.layers.23.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.23.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.23.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.24.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.24.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.24.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.24.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor136                                              (64,)   512.00B
Tensor137                                  (1, 1, 4096, 128)     0.00B
Tensor138                                  (1, 1, 4096, 128)     0.00B
model.layers.24.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.24.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.24.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.25.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.25.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.25.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.25.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor139                                              (64,)   512.00B
Tensor140                                  (1, 1, 4096, 128)     0.00B
Tensor141                                  (1, 1, 4096, 128)     0.00B
model.layers.25.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.25.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.25.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.26.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.26.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.26.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.26.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor142                                              (64,)   512.00B
Tensor143                                  (1, 1, 4096, 128)     0.00B
Tensor144                                  (1, 1, 4096, 128)     0.00B
model.layers.26.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.26.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.26.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.27.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.27.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.27.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.27.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor145                                              (64,)   512.00B
Tensor146                                  (1, 1, 4096, 128)     0.00B
Tensor147                                  (1, 1, 4096, 128)     0.00B
model.layers.27.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.27.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.27.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.28.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.28.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.28.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.28.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor148                                              (64,)   512.00B
Tensor149                                  (1, 1, 4096, 128)     0.00B
```

```
Tensor150                                  (1, 1, 4096, 128)     0.00B
model.layers.28.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.28.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.28.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.29.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.29.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.29.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.29.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor151                                              (64,)   512.00B
Tensor152                                  (1, 1, 4096, 128)     0.00B
Tensor153                                  (1, 1, 4096, 128)     0.00B
model.layers.29.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.29.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.29.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.30.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.30.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.30.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.30.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor154                                              (64,)   512.00B
Tensor155                                  (1, 1, 4096, 128)     0.00B
Tensor156                                  (1, 1, 4096, 128)     0.00B
model.layers.30.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.30.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.30.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.31.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.31.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.31.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.31.self_attn.o_proj.weight         (4096, 4096)    64.00M
Tensor157                                              (64,)   512.00B
Tensor158                                  (1, 1, 4096, 128)     0.00B
Tensor159                                  (1, 1, 4096, 128)     0.00B
model.layers.31.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.31.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.31.mlp.down_proj.weight           (4096, 11008)   172.00M
-------------------------------------------------------------------------------
Total Tensors: 6805526528       Used Memory: 25.23G
-------------------------------------------------------------------------------
```

```
C:\Users\xxc13\anaconda3\Lib\site-packages\pytorch_memlab\mem_reporter.py:65: Fut
ureWarning: `torch.distributed.reduce_op` is deprecated, please use `torch.distri
buted.ReduceOp` instead
  tensors = [obj for obj in objects if isinstance(obj, torch.Tensor)]
C:\Users\xxc13\anaconda3\Lib\site-packages\pytorch_memlab\mem_reporter.py:95: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and Untype
dStorage will be the only storage class. This should only matter to you if you ar
e using storages directly.  To access UntypedStorage directly, use tensor.untyped
_storage() instead of tensor.storage()
  fact_numel = tensor.storage().size()
```

In [ ]: