In [1]:
```
# pip install pytorch_memlab

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

# 模型的snapshot本地地址
model_directory = r"Z:/llmfile/Meta-Llama-3.1-8B-Instruct/models--meta-llama--Me
# model_directory = r"Z:/llmfile/Llama-2-7B-hf/models--meta-llama--Llama-2-7b-hf

from transformers import AutoModelForCausalLM, AutoTokenizer, LlamaConfig

from pytorch_memlab import MemReporter
```

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)
# print(model)
```

In [3]:
```
# 检查一下torch和cuda
# import torch
# print(torch.__version__)
# print(torch.cuda.is_available())
```

In [4]:
```
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%|          | 0/4 [00:00<?, ?it/s]
```

```
Element type                                                 Size  Used MEM
-------------------------------------------------------------------------------
Storage on cpu
lm_head.weight                                (128256, 4096)     1.96G
Tensor0                                               (64,)   512.00B
model.embed_tokens.weight                     (128256, 4096)     1.96G
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
Tensor1                                                      (64,)   512.00B
Tensor2                                                      (64,)   512.00B
Tensor3                                                      (64,)   512.00B
Tensor4                                                      (64,)   512.00B
Tensor5                                                      (64,)   512.00B
Tensor6                                                      (64,)   512.00B
Tensor7                                                      (64,)   512.00B
Tensor8                                                      (64,)   512.00B
Tensor9                                                      (64,)   512.00B
Tensor10                                                     (64,)   512.00B
Tensor11                                                     (64,)   512.00B
Tensor12                                                     (64,)   512.00B
Tensor13                                                     (64,)   512.00B
Tensor14                                                     (64,)   512.00B
Tensor15                                                     (64,)   512.00B
Tensor16                                                     (64,)   512.00B
Tensor17                                                     (64,)   512.00B
Tensor18                                                     (64,)   512.00B
Tensor19                                                     (64,)   512.00B
Tensor20                                                     (64,)   512.00B
Tensor21                                                     (64,)   512.00B
Tensor22                                                     (64,)   512.00B
Tensor23                                                     (64,)   512.00B
Tensor24                                                     (64,)   512.00B
Tensor25                                                     (64,)   512.00B
Tensor26                                                     (64,)   512.00B
Tensor27                                                     (64,)   512.00B
Tensor28                                                     (64,)   512.00B
Tensor29                                                     (64,)   512.00B
Tensor30                                                     (64,)   512.00B
Tensor31                                                     (64,)   512.00B
Tensor32                                                     (64,)   512.00B
model.layers.0.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.0.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.0.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.0.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.0.mlp.gate_proj.weight           (14336, 4096)   224.00M
model.layers.0.mlp.up_proj.weight             (14336, 4096)   224.00M
model.layers.0.mlp.down_proj.weight           (4096, 14336)   224.00M
model.layers.1.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.1.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.1.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.1.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.1.mlp.gate_proj.weight           (14336, 4096)   224.00M
model.layers.1.mlp.up_proj.weight             (14336, 4096)   224.00M
model.layers.1.mlp.down_proj.weight           (4096, 14336)   224.00M
model.layers.2.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.2.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.2.self_attn.v_proj.weight          (1024, 4096)    16.00M
```

```
model.layers.2.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.2.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.2.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.2.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.3.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.3.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.3.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.3.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.3.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.3.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.3.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.4.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.4.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.4.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.4.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.4.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.4.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.4.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.5.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.5.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.5.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.5.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.5.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.5.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.5.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.6.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.6.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.6.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.6.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.6.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.6.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.6.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.7.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.7.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.7.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.7.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.7.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.7.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.7.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.8.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.8.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.8.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.8.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.8.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.8.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.8.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.9.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.9.self_attn.k_proj.weight          (1024, 4096)    16.00M
model.layers.9.self_attn.v_proj.weight          (1024, 4096)    16.00M
model.layers.9.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.9.mlp.gate_proj.weight            (14336, 4096)   224.00M
model.layers.9.mlp.up_proj.weight              (14336, 4096)   224.00M
model.layers.9.mlp.down_proj.weight            (4096, 14336)   224.00M
model.layers.10.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.10.self_attn.k_proj.weight         (1024, 4096)    16.00M
model.layers.10.self_attn.v_proj.weight         (1024, 4096)    16.00M
model.layers.10.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.10.mlp.gate_proj.weight           (14336, 4096)   224.00M
model.layers.10.mlp.up_proj.weight             (14336, 4096)   224.00M
model.layers.10.mlp.down_proj.weight           (4096, 14336)   224.00M
```

```
model.layers.11.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.11.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.11.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.11.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.11.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.11.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.11.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.12.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.12.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.12.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.12.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.12.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.12.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.12.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.13.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.13.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.13.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.13.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.13.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.13.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.13.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.14.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.14.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.14.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.14.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.14.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.14.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.14.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.15.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.15.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.15.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.15.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.15.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.15.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.15.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.16.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.16.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.16.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.16.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.16.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.16.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.16.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.17.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.17.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.17.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.17.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.17.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.17.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.17.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.18.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.18.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.18.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.18.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.18.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.18.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.18.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.19.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.19.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.19.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.19.self_attn.o_proj.weight        (4096, 4096)    64.00M
```

```
model.layers.19.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.19.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.19.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.20.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.20.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.20.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.20.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.20.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.20.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.20.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.21.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.21.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.21.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.21.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.21.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.21.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.21.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.22.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.22.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.22.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.22.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.22.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.22.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.22.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.23.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.23.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.23.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.23.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.23.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.23.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.23.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.24.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.24.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.24.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.24.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.24.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.24.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.24.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.25.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.25.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.25.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.25.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.25.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.25.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.25.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.26.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.26.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.26.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.26.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.26.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.26.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.26.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.27.self_attn.q_proj.weight        (4096, 4096)    64.00M
model.layers.27.self_attn.k_proj.weight        (1024, 4096)    16.00M
model.layers.27.self_attn.v_proj.weight        (1024, 4096)    16.00M
model.layers.27.self_attn.o_proj.weight        (4096, 4096)    64.00M
model.layers.27.mlp.gate_proj.weight          (14336, 4096)   224.00M
model.layers.27.mlp.up_proj.weight            (14336, 4096)   224.00M
model.layers.27.mlp.down_proj.weight          (4096, 14336)   224.00M
model.layers.28.self_attn.q_proj.weight        (4096, 4096)    64.00M
```

```
model.layers.28.self_attn.k_proj.weight         (1024, 4096)    16.00M
model.layers.28.self_attn.v_proj.weight         (1024, 4096)    16.00M
model.layers.28.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.28.mlp.gate_proj.weight           (14336, 4096)   224.00M
model.layers.28.mlp.up_proj.weight             (14336, 4096)   224.00M
model.layers.28.mlp.down_proj.weight           (4096, 14336)   224.00M
model.layers.29.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.29.self_attn.k_proj.weight         (1024, 4096)    16.00M
model.layers.29.self_attn.v_proj.weight         (1024, 4096)    16.00M
model.layers.29.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.29.mlp.gate_proj.weight           (14336, 4096)   224.00M
model.layers.29.mlp.up_proj.weight             (14336, 4096)   224.00M
model.layers.29.mlp.down_proj.weight           (4096, 14336)   224.00M
model.layers.30.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.30.self_attn.k_proj.weight         (1024, 4096)    16.00M
model.layers.30.self_attn.v_proj.weight         (1024, 4096)    16.00M
model.layers.30.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.30.mlp.gate_proj.weight           (14336, 4096)   224.00M
model.layers.30.mlp.up_proj.weight             (14336, 4096)   224.00M
model.layers.30.mlp.down_proj.weight           (4096, 14336)   224.00M
model.layers.31.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.31.self_attn.k_proj.weight         (1024, 4096)    16.00M
model.layers.31.self_attn.v_proj.weight         (1024, 4096)    16.00M
model.layers.31.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.31.mlp.gate_proj.weight           (14336, 4096)   224.00M
model.layers.31.mlp.up_proj.weight             (14336, 4096)   224.00M
model.layers.31.mlp.down_proj.weight           (4096, 14336)   224.00M
--------------------------------------------------------------------------------
Total Tensors: 8030263360       Used Memory: 29.92G
--------------------------------------------------------------------------------
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 [ ]: