```
In [1]: # pip install pytorch_memlab
       # 其他需要的包
       # pip install transformers thop torch
       from transformers import AutoModelForCausalLM, AutoTokenizer, LlamaConfig
       from pytorch_memlab import MemReporter
       import torch
In [2]: # 模型的snapshot本地地址
       # Llama3.1-8b-instruct
       model_directory = r"Z:/llmfile/Meta-Llama-3.1-8B-Instruct/models--meta-llama--Me
       # LLama2-7b
       # model_directory = r"Z:/llmfile/Llama-2-7B-hf/models--meta-llama--Llama-2-7b-hf
In [3]: # 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 [4]: # 检查一下torch和cuda
       # import torch
       # print(torch.__version__)
       # print(torch.cuda.is_available())
In [5]: # 静态的时候的模型内存需求#,输出每层内存使用情况
       # model = AutoModelForCausalLM.from pretrained(model directory)
       # reporter = MemReporter(model)
       # reporter.report()
In [6]: #加载模型和 tokenizer
       model = AutoModelForCausalLM.from_pretrained(model_directory)
       tokenizer = AutoTokenizer.from pretrained(model directory)
       # MemReporter
       reporter = MemReporter(model)
       # 初始输入 token
       input text = "Hello"
       input_ids = tokenizer(input_text, return_tensors="pt").input_ids
       # 推理一个 token 的内存使用情况(前向传播)
       model.eval()
       with torch.no_grad():
           outputs = model(input ids)
           logits = outputs.logits
           reporter.report() # 记录推理过程中的内存使用情况
```

```
# Logits 的形状和内容
print("Logits shape:", logits.shape)
print("Logits:", logits)
```

Loading checkpoint shards: 0% | 0/4 [00:00<?, ?it/s]

Element type Size Used MEM

----------Storage on cpu Tensor0 (1, 2)512.00B Tensor1 (1, 2, 1024)8.00K (1, 8, 2, 128)Tensor2 0.00B Tensor3 (1, 8, 2, 128)8.00K Tensor4 (1, 2, 1024)8.00K (1, 8, 2, 128)Tensor5 0.00B Tensor6 (1, 8, 2, 128)8.00K Tensor7 (1, 2, 1024)8.00K (1, 8, 2, 128) Tensor8 0.00B (1, 8, 2, 128)Tensor9 8.00K Tensor10 (1, 2, 1024)8.00K Tensor11 (1, 8, 2, 128)0.00B Tensor12 (1, 8, 2, 128)8.00K (1, 2, 1024)Tensor13 8.00K (1, 8, 2, 128)Tensor14 0.00B Tensor15 (1, 8, 2, 128)8.00K (1, 2, 1024)Tensor16 8.00K (1, 8, 2, 128)Tensor17 0.00B Tensor18 (1, 8, 2, 128)8.00K (1, 2, 1024)Tensor19 8.00K (1, 8, 2, 128)Tensor20 0.00B Tensor21 (1, 8, 2, 128)8.00K Tensor22 (1, 2, 1024)8.00K (1, 8, 2, 128)Tensor23 0.00B Tensor24 (1, 8, 2, 128)8.00K (1, 2, 1024)Tensor25 8.00K Tensor26 (1, 8, 2, 128)0.00B (1, 8, 2, 128)Tensor27 8.00K Tensor28 (1, 2, 1024)8.00K (1, 8, 2, 128)Tensor29 0.00B (1, 8, 2, 128)Tensor30 8.00K (1, 2, 1024)Tensor31 8.00K Tensor32 (1, 8, 2, 128)0.00B Tensor33 (1, 8, 2, 128)8.00K (1, 2, 1024)Tensor34 8.00K (1, 8, 2, 128)Tensor35 0.00B (1, 8, 2, 128)Tensor36 8.00K (1, 2, 1024)Tensor37 8.00K (1, 8, 2, 128)Tensor38 0.00B (1, 8, 2, 128)Tensor39 8.00K Tensor40 (1, 2, 1024)8.00K Tensor41 (1, 8, 2, 128)0.00B (1, 8, 2, 128)Tensor42 8.00K Tensor43 (1, 2, 1024)8.00K Tensor44 (1, 8, 2, 128)0.00B Tensor45 (1, 8, 2, 128)8.00K (1, 2, 1024)Tensor46 8.00K (1, 8, 2, 128)Tensor47 0.00B (1, 8, 2, 128) Tensor48 8.00K (1, 2, 1024)Tensor49 8.00K (1, 8, 2, 128)Tensor50 0.00B Tensor51 (1, 8, 2, 128)8.00K Tensor52 (1, 2, 1024)8.00K (1, 8, 2, 128)Tensor53 0.00B Tensor54 (1, 8, 2, 128)8.00K Tensor55 (1, 2, 1024)8.00K Tensor56 (1, 8, 2, 128)0.00B

```
Tensor57
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor58
                                                   (1, 2, 1024)
                                                                    8.00K
Tensor59
                                                (1, 8, 2, 128)
                                                                    0.00B
Tensor60
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor61
                                                   (1, 2, 1024)
                                                                    8.00K
Tensor62
                                                (1, 8, 2, 128)
                                                                    0.00B
Tensor63
                                                (1, 8, 2, 128)
                                                                    8.00K
                                                   (1, 2, 1024)
Tensor64
                                                                    8.00K
                                                (1, 8, 2, 128)
Tensor65
                                                                    0.00B
Tensor66
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor67
                                                   (1, 2, 1024)
                                                                    8.00K
Tensor68
                                                (1, 8, 2, 128)
                                                                    0.00B
                                                (1, 8, 2, 128)
Tensor69
                                                                    8.00K
Tensor70
                                                   (1, 2, 1024)
                                                                    8.00K
Tensor71
                                                (1, 8, 2, 128)
                                                                    0.00B
Tensor72
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor73
                                                   (1, 2, 1024)
                                                                    8.00K
                                                (1, 8, 2, 128)
Tensor74
                                                                    0.00B
Tensor75
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor76
                                                   (1, 2, 1024)
                                                                    8.00K
                                                (1, 8, 2, 128)
Tensor77
                                                                    0.00B
Tensor78
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor79
                                                   (1, 2, 1024)
                                                                    8.00K
                                                (1, 8, 2, 128)
Tensor80
                                                                    0.00B
Tensor81
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor82
                                                   (1, 2, 1024)
                                                                    8.00K
                                                (1, 8, 2, 128)
Tensor83
                                                                    0.00B
Tensor84
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor85
                                                   (1, 2, 1024)
                                                                    8.00K
Tensor86
                                                (1, 8, 2, 128)
                                                                    0.00B
Tensor87
                                                (1, 8, 2, 128)
                                                                    8.00K
Tensor88
                                                   (1, 2, 1024)
                                                                    8.00K
                                                                    0.00B
Tensor89
                                                (1, 8, 2, 128)
Tensor90
                                                (1, 8, 2, 128)
                                                                    8.00K
                                                   (1, 2, 1024)
Tensor91
                                                                    8.00K
Tensor92
                                                (1, 8, 2, 128)
                                                                    0.00B
Tensor93
                                                (1, 8, 2, 128)
                                                                    8.00K
                                                   (1, 2, 1024)
Tensor94
                                                                    8.00K
Tensor95
                                                (1, 8, 2, 128)
                                                                    0.00B
                                                (1, 8, 2, 128)
Tensor96
                                                                    8.00K
Tensor97
                                                (1, 2, 128256)
                                                                 1002.00K
lm head.weight
                                                (128256, 4096)
                                                                    1.96G
Tensor98
                                                          (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
                                                              (4096,)
model.layers.1.post_attention_layernorm.weight
                                                                          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
                                                                   16.00K
model.layers.4.input layernorm.weight
                                                        (4096,)
model.layers.4.post_attention_layernorm.weight
                                                              (4096,)
                                                                          16.00K
                                                        (4096,)
model.layers.5.input_layernorm.weight
                                                                   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
                                                                   16.00K
                                                        (4096,)
```

```
model.layers.7.post_attention_layernorm.weight
                                                             (4096,)
                                                                         16.00K
                                                       (4096,)
model.layers.8.input_layernorm.weight
                                                                  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
                                                                  16.00K
                                                       (4096,)
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
                                                              (4096,)
model.layers.18.post_attention_layernorm.weight
                                                                         16.00K
                                                       (4096,)
model.layers.19.input_layernorm.weight
                                                                  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
                                                                  16.00K
model.layers.25.input layernorm.weight
                                                       (4096,)
model.layers.25.post attention layernorm.weight
                                                              (4096,)
                                                                         16.00K
model.layers.26.input_layernorm.weight
                                                       (4096,)
                                                                  16.00K
                                                              (4096,)
model.layers.26.post attention layernorm.weight
                                                                          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
                                                       (4096,)
model.layers.29.input_layernorm.weight
                                                                  16.00K
model.layers.29.post_attention_layernorm.weight
                                                              (4096,)
                                                                         16.00K
                                                       (4096,)
model.layers.30.input_layernorm.weight
                                                                  16.00K
model.layers.30.post_attention_layernorm.weight
                                                              (4096,)
                                                                          16.00K
                                                       (4096,)
model.layers.31.input layernorm.weight
                                                                  16.00K
model.layers.31.post_attention_layernorm.weight
                                                              (4096,)
                                                                         16.00K
                                                         (64,)
Tensor99
                                                                 512.00B
Tensor100
                                                         (64,)
                                                                 512.00B
Tensor101
                                                         (64,)
                                                                 512.00B
                                                                 512.00B
Tensor102
                                                         (64,)
Tensor103
                                                         (64,)
                                                                 512.00B
Tensor104
                                                         (64,)
                                                                 512.00B
Tensor105
                                                         (64,)
                                                                 512.00B
Tensor106
                                                         (64,)
                                                                 512.00B
Tensor107
                                                         (64,)
                                                                 512.00B
Tensor108
                                                         (64,)
                                                                 512.00B
                                                                 512.00B
Tensor109
                                                         (64,)
```

Tensor110	(64,)	512.00B
Tensor111	(64,)	512.00B
Tensor112	(64,)	512.00B
Tensor113	(64,)	512.00B
Tensor114	(64,)	512.00B
Tensor115	(64,)	512.00B
Tensor116	(64,)	512.00B
Tensor117	(64,)	512.00B
Tensor118	(64,)	512.00B
Tensor119	(64,)	512.00B
Tensor120	(64,)	512.00B
Tensor121	(64,)	512.00B
Tensor122	(64,)	
Tensor123	(64,)	
Tensor124	(64,)	
Tensor125	(64,)	
Tensor126	(64,)	512.00B
Tensor127	(64,)	512.00B
Tensor128	(64,)	512.00B
Tensor129	(64,)	512.00B
Tensor130	(64,)	
	(4096, 4096)	
model.layers.0.self_attn.q_proj.weight	, , ,	
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
<pre>model.layers.1.self_attn.o_proj.weight</pre>	(4096, 4096)	64.00M
<pre>model.layers.1.mlp.gate_proj.weight</pre>	(14336, 4096)	224.00M
<pre>model.layers.1.mlp.up_proj.weight</pre>	(14336, 4096)	224.00M
<pre>model.layers.1.mlp.down_proj.weight</pre>	(4096, 14336)	224.00M
<pre>model.layers.2.self_attn.q_proj.weight</pre>	(4096, 4096)	64.00M
<pre>model.layers.2.self_attn.k_proj.weight</pre>	(1024, 4096)	16.00M
<pre>model.layers.2.self_attn.v_proj.weight</pre>	(1024, 4096)	16.00M
<pre>model.layers.2.self_attn.o_proj.weight</pre>	(4096, 4096)	64.00M
<pre>model.layers.2.mlp.gate_proj.weight</pre>	(14336, 4096)	224.00M
<pre>model.layers.2.mlp.up_proj.weight</pre>	(14336, 4096)	224.00M
<pre>model.layers.2.mlp.down_proj.weight</pre>	(4096, 14336)	224.00M
<pre>model.layers.3.self_attn.q_proj.weight</pre>	(4096, 4096)	64.00M
<pre>model.layers.3.self_attn.k_proj.weight</pre>	(1024, 4096)	16.00M
<pre>model.layers.3.self_attn.v_proj.weight</pre>	(1024, 4096)	16.00M
<pre>model.layers.3.self_attn.o_proj.weight</pre>	(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
moder.rayer 3.3.3err accii.o_proj.werBill	(4030, 4030)	04.0011

```
model.layers.5.mlp.gate_proj.weight
                                                (14336, 4096)
                                                                 224.00M
                                                (14336, 4096)
model.layers.5.mlp.up_proj.weight
                                                                 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
                                                (14336, 4096)
model.layers.6.mlp.up_proj.weight
                                                                 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
                                                 (1024, 4096)
model.layers.7.self_attn.k_proj.weight
                                                                  16.00M
                                                 (1024, 4096)
model.layers.7.self_attn.v_proj.weight
                                                                  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
                                                 (4096, 4096)
model.layers.8.self_attn.o_proj.weight
                                                                  64.00M
model.layers.8.mlp.gate_proj.weight
                                                (14336, 4096)
                                                                 224.00M
model.layers.8.mlp.up_proj.weight
                                                (14336, 4096)
                                                                 224.00M
                                                (4096, 14336)
model.layers.8.mlp.down_proj.weight
                                                                 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
                                                (4096, 14336)
model.layers.9.mlp.down_proj.weight
                                                                 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
                                                 (1024, 4096)
model.layers.10.self_attn.v_proj.weight
                                                                  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
                                                (4096, 14336)
model.layers.10.mlp.down_proj.weight
                                                                 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
                                                 (4096, 4096)
model.layers.12.self_attn.o_proj.weight
                                                                  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
                                                 (1024, 4096)
model.layers.13.self_attn.k_proj.weight
                                                                  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
                                                (14336, 4096)
model.layers.13.mlp.gate_proj.weight
                                                                 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
                                                 (4096, 14336)
model.layers.15.mlp.down_proj.weight
                                                                 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
                                                 (4096, 4096)
model.layers.17.self_attn.o_proj.weight
                                                                  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
                                                 (4096, 4096)
model.layers.18.self_attn.o_proj.weight
                                                                  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
                                                 (14336, 4096)
model.layers.20.mlp.gate_proj.weight
                                                                 224.00M
model.layers.20.mlp.up proj.weight
                                                 (14336, 4096)
                                                                 224.00M
model.layers.20.mlp.down proj.weight
                                                 (4096, 14336)
                                                                 224.00M
                                                 (4096, 4096)
model.layers.21.self_attn.q_proj.weight
                                                                  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
                                                 (14336, 4096)
model.layers.21.mlp.up proj.weight
                                                                 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
                                                 (1024, 4096)
model.layers.22.self_attn.k_proj.weight
                                                                  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
                                                 (1024, 4096)
model.layers.23.self_attn.v_proj.weight
                                                                  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
                                                (4096, 14336)
model.layers.23.mlp.down_proj.weight
                                                                 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
                                                 (1024, 4096)
model.layers.24.self_attn.v_proj.weight
                                                                  16.00M
                                                 (4096, 4096)
model.layers.24.self_attn.o_proj.weight
                                                                  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
                                                (14336, 4096)
model.layers.28.mlp.up_proj.weight
                                                                 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
                                                (14336, 4096)
model.layers.29.mlp.gate_proj.weight
                                                                 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
                                                 (1024, 4096)
model.layers.30.self_attn.v_proj.weight
                                                                  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)
                                         (4096, 14336)
model.layers.31.mlp.down_proj.weight
                                                        224.00M
-----
Total Tensors: 8030716482
                           Used Memory: 29.92G
______
Logits shape: torch.Size([1, 2, 128256])
Logits: tensor([[[-2.9512, 1.5942, 7.4129, ..., 1.6402, 1.6403, 1.6403],
        [14.2607, 2.0137, 2.5573, \ldots, -2.9703, -2.9704, -2.9706]]])
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 []: