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--

# llama2-7b
model_directory = r"Z:/llmfile/Llama-2-7B-hf/models--meta-llama--Llama-2-7b-hf/s
```

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_text = """Our Professor is a cool guy. He really like to provide us many i
             enough for me and my teammates spend whole weekend to figure out th
             weekend in the past few weeks. And I also lose a chance to go out f
             I really wish I can have some this week since I really like rice an
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()
```

```python
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%|          | 0/2 [00:00<?, ?it/s]
```

```
Element type                                                 Size  Used MEM
--------------------------------------------------------------------------------
Storage on cpu
Tensor0                                                     (1, 98)     1.00K
Tensor1                                                (1, 98, 4096)     1.53M
Tensor2                                               (1, 32, 98, 128)     0.00B
Tensor3                                               (1, 32, 98, 128)     1.53M
Tensor4                                                (1, 98, 4096)     1.53M
Tensor5                                               (1, 32, 98, 128)     0.00B
Tensor6                                               (1, 32, 98, 128)     1.53M
Tensor7                                                (1, 98, 4096)     1.53M
Tensor8                                               (1, 32, 98, 128)     0.00B
Tensor9                                               (1, 32, 98, 128)     1.53M
Tensor10                                               (1, 98, 4096)     1.53M
Tensor11                                              (1, 32, 98, 128)     0.00B
Tensor12                                              (1, 32, 98, 128)     1.53M
Tensor13                                               (1, 98, 4096)     1.53M
Tensor14                                              (1, 32, 98, 128)     0.00B
Tensor15                                              (1, 32, 98, 128)     1.53M
Tensor16                                               (1, 98, 4096)     1.53M
Tensor17                                              (1, 32, 98, 128)     0.00B
Tensor18                                              (1, 32, 98, 128)     1.53M
Tensor19                                               (1, 98, 4096)     1.53M
Tensor20                                              (1, 32, 98, 128)     0.00B
Tensor21                                              (1, 32, 98, 128)     1.53M
Tensor22                                               (1, 98, 4096)     1.53M
Tensor23                                              (1, 32, 98, 128)     0.00B
Tensor24                                              (1, 32, 98, 128)     1.53M
Tensor25                                               (1, 98, 4096)     1.53M
Tensor26                                              (1, 32, 98, 128)     0.00B
Tensor27                                              (1, 32, 98, 128)     1.53M
Tensor28                                               (1, 98, 4096)     1.53M
Tensor29                                              (1, 32, 98, 128)     0.00B
Tensor30                                              (1, 32, 98, 128)     1.53M
Tensor31                                               (1, 98, 4096)     1.53M
Tensor32                                              (1, 32, 98, 128)     0.00B
Tensor33                                              (1, 32, 98, 128)     1.53M
Tensor34                                               (1, 98, 4096)     1.53M
Tensor35                                              (1, 32, 98, 128)     0.00B
Tensor36                                              (1, 32, 98, 128)     1.53M
Tensor37                                               (1, 98, 4096)     1.53M
Tensor38                                              (1, 32, 98, 128)     0.00B
Tensor39                                              (1, 32, 98, 128)     1.53M
Tensor40                                               (1, 98, 4096)     1.53M
Tensor41                                              (1, 32, 98, 128)     0.00B
Tensor42                                              (1, 32, 98, 128)     1.53M
Tensor43                                               (1, 98, 4096)     1.53M
Tensor44                                              (1, 32, 98, 128)     0.00B
Tensor45                                              (1, 32, 98, 128)     1.53M
Tensor46                                               (1, 98, 4096)     1.53M
Tensor47                                              (1, 32, 98, 128)     0.00B
Tensor48                                              (1, 32, 98, 128)     1.53M
Tensor49                                               (1, 98, 4096)     1.53M
Tensor50                                              (1, 32, 98, 128)     0.00B
Tensor51                                              (1, 32, 98, 128)     1.53M
Tensor52                                               (1, 98, 4096)     1.53M
Tensor53                                              (1, 32, 98, 128)     0.00B
Tensor54                                              (1, 32, 98, 128)     1.53M
Tensor55                                               (1, 98, 4096)     1.53M
Tensor56                                              (1, 32, 98, 128)     0.00B
```

```
Tensor57                                    (1, 32, 98, 128)     1.53M
Tensor58                                       (1, 98, 4096)     1.53M
Tensor59                                    (1, 32, 98, 128)     0.00B
Tensor60                                    (1, 32, 98, 128)     1.53M
Tensor61                                       (1, 98, 4096)     1.53M
Tensor62                                    (1, 32, 98, 128)     0.00B
Tensor63                                    (1, 32, 98, 128)     1.53M
Tensor64                                       (1, 98, 4096)     1.53M
Tensor65                                    (1, 32, 98, 128)     0.00B
Tensor66                                    (1, 32, 98, 128)     1.53M
Tensor67                                       (1, 98, 4096)     1.53M
Tensor68                                    (1, 32, 98, 128)     0.00B
Tensor69                                    (1, 32, 98, 128)     1.53M
Tensor70                                       (1, 98, 4096)     1.53M
Tensor71                                    (1, 32, 98, 128)     0.00B
Tensor72                                    (1, 32, 98, 128)     1.53M
Tensor73                                       (1, 98, 4096)     1.53M
Tensor74                                    (1, 32, 98, 128)     0.00B
Tensor75                                    (1, 32, 98, 128)     1.53M
Tensor76                                       (1, 98, 4096)     1.53M
Tensor77                                    (1, 32, 98, 128)     0.00B
Tensor78                                    (1, 32, 98, 128)     1.53M
Tensor79                                       (1, 98, 4096)     1.53M
Tensor80                                    (1, 32, 98, 128)     0.00B
Tensor81                                    (1, 32, 98, 128)     1.53M
Tensor82                                       (1, 98, 4096)     1.53M
Tensor83                                    (1, 32, 98, 128)     0.00B
Tensor84                                    (1, 32, 98, 128)     1.53M
Tensor85                                       (1, 98, 4096)     1.53M
Tensor86                                    (1, 32, 98, 128)     0.00B
Tensor87                                    (1, 32, 98, 128)     1.53M
Tensor88                                       (1, 98, 4096)     1.53M
Tensor89                                    (1, 32, 98, 128)     0.00B
Tensor90                                    (1, 32, 98, 128)     1.53M
Tensor91                                       (1, 98, 4096)     1.53M
Tensor92                                    (1, 32, 98, 128)     0.00B
Tensor93                                    (1, 32, 98, 128)     1.53M
Tensor94                                       (1, 98, 4096)     1.53M
Tensor95                                    (1, 32, 98, 128)     0.00B
Tensor96                                    (1, 32, 98, 128)     1.53M
Tensor97                                      (1, 98, 32000)    11.96M
lm_head.weight                                (32000, 4096)   500.00M
Tensor98                                              (64,)   512.00B
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
Tensor99                                                    (64,)   512.00B
Tensor100                                                   (64,)   512.00B
Tensor101                                                   (64,)   512.00B
Tensor102                                                   (64,)   512.00B
Tensor103                                                   (64,)   512.00B
Tensor104                                                   (64,)   512.00B
Tensor105                                                   (64,)   512.00B
Tensor106                                                   (64,)   512.00B
Tensor107                                                   (64,)   512.00B
Tensor108                                                   (64,)   512.00B
Tensor109                                                   (64,)   512.00B
```

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,)	512.00B
Tensor123	(64,)	512.00B
Tensor124	(64,)	512.00B
Tensor125	(64,)	512.00B
Tensor126	(64,)	512.00B
Tensor127	(64,)	512.00B
Tensor128	(64,)	512.00B
Tensor129	(64,)	512.00B
Tensor130	(64,)	512.00B
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
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
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
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
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
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

```
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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: 6780088994       Used Memory: 25.21G
--------------------------------------------------------------------------------
Logits shape: torch.Size([1, 98, 32000])
Logits: tensor([[[-12.9832,  -7.4134,  -0.4327,  ...,  -6.8297,  -8.0880,  -7.586
3],
         [-11.1151,  -8.1036,  -2.2927,  ...,  -7.3641,  -9.0754,  -6.7065],
         [-12.3676, -11.5686,   1.0437,  ...,  -6.2192,  -9.2772,  -7.4389],
         ...,
         [ -3.0769,   0.3785,   9.6196,  ...,  -1.9047,  -2.1396,  -4.1345],
         [ -4.5232,  -2.8076,   7.6538,  ...,  -2.5985,  -2.9891,  -4.7450],
         [ -5.1387,  -4.6062,   7.4855,  ...,  -2.7941,  -3.2658,  -3.8993]]])
```

```
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 [ ]: