```
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()
       # 加载模型和 tokenizer
In [3]:
       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%| | 0/2 [00:00<?, ?it/s]

Element type Size Used MEM

Tensor1 Tensor2 Tensor3 Tensor3 Tensor3 Tensor3 Tensor4 Tensor5 Tensor5 Tensor6 Tensor6 Tensor7 Tensor7 Tensor7 Tensor7 Tensor7 Tensor8 Tensor8 Tensor9 Tensor9 Tensor10 Tensor1 Tensor11 Tensor11 Tensor11 Tensor12 Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor19 Tensor19 Tensor19 Tensor10 Tensor10 Tensor10 Tensor10 Tensor11 Tensor11 Tensor11 Tensor12 Tensor13 Tensor14 Tensor15 Tensor16 Tensor17 Tensor17 Tensor19 Tensor19 Tensor19 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor22 Tensor22 Tensor23 Tensor24 Tensor25 Tensor26 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor30 Tensor	
Tensor1 Tensor2 Tensor3 Tensor3 Tensor3 Tensor3 Tensor4 Tensor5 Tensor5 Tensor6 Tensor6 Tensor7 Tensor7 Tensor7 Tensor7 Tensor7 Tensor7 Tensor7 Tensor8 Tensor8 Tensor9 Tensor9 Tensor10 Tensor11 Tensor11 Tensor11 Tensor12 Tensor12 Tensor13 Tensor14 Tensor14 Tensor15 Tensor15 Tensor16 Tensor17 Tensor17 Tensor17 Tensor17 Tensor18 Tensor19 Tensor19 Tensor19 Tensor10 Tensor10 Tensor10 Tensor10 Tensor11 Tensor11 Tensor11 Tensor11 Tensor12 Tensor12 Tensor13 Tensor14 Tensor15 Tensor16 Tensor17 Tensor17 Tensor17 Tensor19 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor29 Tensor29 Tensor29 Tensor30 Tensor30 Tensor31 Tensor30 Tensor31 Tensor30 Tensor31 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor39 Tensor37 Tensor39 Tensor39 Tensor39 Tensor30 Tensor3	12.00B
Tensor2 Tensor3 Tensor3 Tensor4 Tensor5 Tensor6 Tensor6 Tensor7 Tensor7 Tensor7 Tensor7 Tensor8 Tensor8 Tensor9 Tensor9 Tensor9 Tensor9 Tensor10 Tensor11 Tensor11 Tensor12 Tensor12 Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor15 Tensor16 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor10 Tensor10 Tensor10 Tensor10 Tensor10 Tensor11 Tensor11 Tensor12 Tensor12 Tensor13 Tensor15 Tensor16 Tensor17 Tensor17 Tensor18 Tensor19 Tensor20 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor21 Tensor22 Tensor22 Tensor23 Tensor24 Tensor25 Tensor25 Tensor27 Tensor28 Tensor29 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor32 Tensor31 Tensor31 Tensor32 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor39 Tensor36 Tensor37 Tensor37 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor32 Tensor33 Tensor34 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor30 Te	32.00K
Tensor3 Tensor4 Tensor5 Tensor6 Tensor6 Tensor6 Tensor7 Tensor7 Tensor7 Tensor7 Tensor8 Tensor8 Tensor8 Tensor9 Tensor9 Tensor9 Tensor10 Tensor11 Tensor12 Tensor13 Tensor13 Tensor14 Tensor15 Tensor15 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor19 Tensor10 Tensor10 Tensor10 Tensor11 Tensor10 Tensor11 Tensor11 Tensor11 Tensor12 Tensor13 Tensor15 Tensor16 Tensor17 Tensor17 Tensor18 Tensor19 Tensor19 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor22 Tensor23 Tensor24 Tensor25 Tensor27 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor30 T	0.00B
Tensor4 Tensor5 Tensor6 Tensor6 (1, 32, 2, 128) Tensor7 (1, 2, 4096) Tensor8 Tensor8 (1, 32, 2, 128) Tensor9 Tensor10 Tensor11 (1, 32, 2, 128) Tensor13 Tensor13 Tensor14 Tensor15 Tensor15 Tensor16 (1, 32, 2, 128) Tensor16 Tensor17 Tensor17 Tensor18 Tensor19 Tensor19 Tensor10 Tensor17 Tensor18 (1, 32, 2, 128) Tensor18 (1, 32, 2, 128) Tensor19 Tensor19 Tensor19 Tensor20 (1, 32, 2, 128) Tensor21 Tensor21 Tensor21 Tensor22 (1, 2, 4096) Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor30 Tensor30 Tensor31 Tensor31 Tensor31 Tensor32 Tensor32 Tensor33 Tensor34 Tensor34 Tensor35 Tensor36 Tensor37 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor39 Tensor40 Tensor50 Tensor40 Tensor50	32.00K
Tensor5 Tensor6 Tensor6 Tensor7 Tensor7 Tensor8 Tensor9 Tensor9 Tensor10 Tensor11 Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor19 Tensor10 Tensor10 Tensor10 Tensor10 Tensor11 Tensor11 Tensor12 Tensor12 Tensor13 Tensor14 Tensor15 Tensor16 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor24 Tensor25 Tensor26 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor39 Tensor31 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor30 Tensor31 Tensor30 Tensor31 Tensor30 Tensor31 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor39 Tensor39 Tensor30 Tensor40 Tensor50 Tensor40 Tensor50 Tensor40 Tensor50 Tensor40 Tensor50 Te	32.00K
Tensor6 Tensor7 Tensor8 Tensor9 Tensor9 Tensor10 Tensor10 Tensor11 Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor18 Tensor18 Tensor18 Tensor19 Tensor11 Tensor11 Tensor11 Tensor12 Tensor13 Tensor14 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor19 Tensor19 Tensor19 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor23 Tensor24 Tensor25 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor39 Tensor30 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor40 Tensor50 Tensor50 Tensor50 Tensor51 Tensor51 Tensor52 Tensor52 Tensor53 Tensor53 Tensor53 Tensor55 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor5	0.00B
Tensor7 Tensor8 (1, 2, 4096) Tensor8 (1, 32, 2, 128) Tensor9 Tensor10 Tensor11 (1, 32, 2, 128) Tensor12 Tensor13 Tensor13 Tensor14 Tensor15 Tensor15 Tensor15 Tensor16 (1, 32, 2, 128) Tensor17 Tensor17 Tensor17 Tensor18 Tensor18 Tensor18 Tensor19 Tensor19 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor28 Tensor29 Tensor39 Tensor29 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor40 Tensor50 Tensor40 Tensor50	32.00K
Tensor8 Tensor9 Tensor10 Tensor10 Tensor11 Tensor11 Tensor12 Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor19 Tensor19 Tensor10 Tensor10 Tensor11 Tensor11 Tensor11 Tensor11 Tensor11 Tensor15 Tensor16 Tensor17 Tensor17 Tensor18 Tensor19 Tensor19 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor24 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor32 Tensor32 Tensor33 Tensor34 Tensor35 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor40 Tensor50 Tenso	32.00K
Tensor9 Tensor10 Tensor11 Tensor12 Tensor12 Tensor13 Tensor13 Tensor14 Tensor15 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor17 Tensor18 Tensor19 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor21 Tensor21 Tensor22 Tensor33 Tensor24 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor34 Tensor35 Tensor35 Tensor36 Tensor36 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor30 Tensor40 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor51 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor54 Tensor54 Tensor55 Tensor55 Tensor55 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor54 Tensor55 Tensor54 Tensor54 Tensor55 Tensor55 Tensor54 Tensor55 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tens	0.00B
Tensor10 Tensor11 Tensor12 Tensor12 Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor15 Tensor16 Tensor16 Tensor16 Tensor17 Tensor17 Tensor17 Tensor17 Tensor18 Tensor19 Tensor19 Tensor19 Tensor19 Tensor19 Tensor19 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor30 Tensor20 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor31 Tensor33 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor40 Tensor50 Ten	32.00K
Tensor11 Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor23 Tensor24 Tensor25 Tensor26 Tensor27 Tensor27 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor4 Tensor30 Tensor4 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor39 Tensor40 Tensor39 Tensor40 Tensor39 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor45 Tensor45 Tensor47 Tensor48 Tensor49 Tensor50 Tenso	32.00K
Tensor12 Tensor13 Tensor14 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor21 Tensor22 Tensor23 Tensor23 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor30 Tensor31 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor40 Tensor50 Ten	0.00B
Tensor13 Tensor14 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor19 Tensor19 Tensor19 Tensor20 Tensor20 Tensor21 Tensor22 Tensor23 Tensor25 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor29 Tensor39 Tensor30 Tensor4 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor40 Tensor40 Tensor40 Tensor41 Tensor42 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor47 Tensor48 Tensor40 Tensor50 Tens	32.00K
Tensor14 Tensor15 Tensor15 Tensor16 Tensor16 Tensor17 Tensor17 Tensor18 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor22 Tensor23 Tensor24 Tensor25 Tensor26 Tensor26 Tensor27 Tensor38 Tensor27 Tensor38 Tensor39 Tensor40 Tensor39 Tensor30 Tensor30 Tensor30 Tensor40 Tensor40 Tensor40 Tensor40 Tensor40 Tensor40 Tensor51 Tensor51 Tensor60 Tensor60 Tensor70 Tensor40 Ten	32.00K
Tensor15 Tensor16 Tensor17 Tensor17 Tensor18 Tensor18 Tensor19 Tensor20 Tensor21 Tensor21 Tensor22 Tensor23 Tensor23 Tensor25 Tensor26 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor20 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor22 Tensor23 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor38 Tensor38 Tensor39 Tensor40 Tensor50 Ten	0.00B
Tensor16 Tensor17 Tensor18 Tensor18 Tensor19 Tensor19 Tensor20 Tensor21 Tensor21 Tensor22 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor21 Tensor21 Tensor22 Tensor23 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor38 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor38 Tensor39 Tensor39 Tensor40 Tensor50 Ten	32.00K
Tensor17 Tensor18 Tensor18 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor22 Tensor23 Tensor23 Tensor25 Tensor25 Tensor26 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor21 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor20 Tensor30 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor32 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor38 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor40 Tensor39 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor42 Tensor41 Tensor42 Tensor40 Tensor41 Tensor41 Tensor41 Tensor41 Tensor41 Tensor42 Tensor41 Tensor51 Tensor52 Tensor52 Tensor53 Tensor53 Tensor53 Tensor53 Tensor53 Tensor53 Tensor54 Tensor53 Tensor54 Tensor55 Tensor54 Tensor54 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor56 Tensor57 Tensor56 Tensor57 Tensor57 Tensor59 Tensor59 Tensor59 Tensor59 Tensor59 Tensor50 Tensor54 Tensor59 Tensor54 Tensor59 Tensor54 Tensor59 Tensor54 Tensor59 Tensor54 Tensor59 Ten	32.00K
Tensor18 Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor22 Tensor22 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor28 Tensor28 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor41 Tensor30 Tensor31 Tensor32 Tensor32 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor38 Tensor39 Tensor40 Tensor39 Tensor40 Tensor40 Tensor40 Tensor41 Tensor42 Tensor42 Tensor45 Tensor45 Tensor46 Tensor47 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor49 Tensor40 Tensor50 Ten	0.00B
Tensor19 Tensor20 Tensor20 Tensor21 Tensor21 Tensor22 Tensor22 Tensor23 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor32 Tensor32 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor38 Tensor39 Tensor40 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor51 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor54 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor59 Tensor59 Tensor50 Ten	32.00K
Tensor20 Tensor21 Tensor21 Tensor22 Tensor22 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor28 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor32 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor38 Tensor39 Tensor40 Tensor40 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor40 Tensor47 Tensor41 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor45 Tensor49 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor45 Tensor45 Tensor49 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor45 Tensor45 Tensor49 Tensor40 Tensor40 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor53 Tensor54 Tensor53 Tensor55 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor56 Tensor56 Tensor56 Tensor57 Tensor56 Tensor57 Tensor58 Tensor59 Ten	32.00K
Tensor21 Tensor22 Tensor22 Tensor23 Tensor23 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor28 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor32 Tensor32 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor36 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor40 Tensor37 Tensor38 Tensor40 Tensor40 Tensor41 Tensor42 Tensor42 Tensor42 Tensor43 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor40 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor40 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor40 Tensor40 Tensor40 Tensor41 Tensor41 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor48 Tensor49 Tensor49 Tensor50 Tensor40 Tensor50 Ten	0.00B
Tensor22 Tensor23 Tensor23 Tensor24 Tensor25 Tensor26 Tensor27 Tensor27 Tensor27 Tensor28 Tensor28 Tensor29 Tensor29 Tensor30 Tensor30 Tensor30 Tensor30 Tensor31 Tensor31 Tensor32 Tensor32 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor38 Tensor38 Tensor39 Tensor39 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor40 Tensor50 Tensor40 Tensor50 Tensor41 Tensor50 Ten	32.00K
Tensor23 Tensor24 Tensor24 Tensor25 Tensor25 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor30 Tensor30 Tensor31 Tensor32 Tensor32 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor40 Tensor40 Tensor41 Tensor42 Tensor42 Tensor42 Tensor45 Tensor45 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor42 Tensor43 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor40 Tensor41 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor49 Tensor49 Tensor49 Tensor49 Tensor49 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor54 Tensor55 Tensor5 Tensor5 Tensor5 Tensor5 Tensor5 Tensor5 Tensor5 Tensor5 Tensor5 Ten	32.00K
Tensor24 Tensor25 Tensor26 Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor29 Tensor30 Tensor30 Tensor31 Tensor32 Tensor32 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor40 Tensor40 Tensor41 Tensor42 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor49 Tensor49 Tensor49 Tensor50 Ten	0.00B
Tensor 25 (1, 2, 4096) Tensor 26 (1, 32, 2, 128) Tensor 27 (1, 32, 2, 128) Tensor 28 (1, 2, 4096) Tensor 29 (1, 32, 2, 128) Tensor 30 (1, 32, 2, 128) Tensor 31 (1, 2, 4096) Tensor 32 (1, 32, 2, 128) Tensor 33 (1, 32, 2, 128) Tensor 34 (1, 2, 4096) Tensor 35 (1, 32, 2, 128) Tensor 36 (1, 32, 2, 128) Tensor 37 (1, 2, 4096) Tensor 38 (1, 32, 2, 128) Tensor 39 (1, 32, 2, 128) Tensor 40 (1, 32, 2, 128) Tensor 41 (1, 2, 4096) Tensor 42 (1, 32, 2, 128) Tensor 43 (1, 2, 4096) Tensor 44 (1, 32, 2, 128) Tensor 45 (1, 32, 2, 128) Tensor 47 (1, 32, 2, 128) Tensor 48 (1, 32, 2, 128) Tensor 49 (1, 32, 2, 128) Tensor 50 (1, 32, 2, 128) Tensor 51 (1, 32, 2, 128) Tensor 52 (1, 2, 4096) Tensor 53	32.00K
Tensor26 Tensor27 Tensor27 Tensor28 Tensor29 Tensor30 Tensor31 Tensor32 Tensor32 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor39 Tensor39 Tensor40 Tensor41 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor44 Tensor45 Tensor45 Tensor47 Tensor47 Tensor48 Tensor48 Tensor49 Tensor50 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor58 Tensor59 Ten	32.00K
Tensor27 Tensor28 Tensor29 Tensor30 Tensor31 Tensor32 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor38 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor39 Tensor40 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor49 Tensor40 Tensor41 Tensor41 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor48 Tensor49 Tensor59 Tensor50 Tensor50 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor58 Tensor59 Ten	0.00B
Tensor 28 Tensor 29 Tensor 29 Tensor 30 Tensor 30 Tensor 31 Tensor 32 Tensor 32 Tensor 33 Tensor 34 Tensor 35 Tensor 35 Tensor 36 Tensor 37 Tensor 37 Tensor 38 Tensor 38 Tensor 39 Tensor 39 Tensor 39 Tensor 40 Tensor 41 Tensor 42 Tensor 42 Tensor 43 Tensor 44 Tensor 44 Tensor 45 Tensor 45 Tensor 47 Tensor 48 Tensor 48 Tensor 49 Tensor 49 Tensor 49 Tensor 49 Tensor 49 Tensor 49 Tensor 40 Tensor 40 Tensor 41 Tensor 42 Tensor 43 Tensor 45 Tensor 46 Tensor 47 Tensor 48 Tensor 49 Tensor 50 Tensor 50 Tensor 50 Tensor 51 Tensor 52 Tensor 53 Tensor 53 Tensor 54 Tensor 55 Tensor 55 Tensor 56 Tensor 57 Tensor 57 Tensor 57 Tensor 58 Tensor 59 Tensor 50 Tensor 50 Tensor 50 Tensor 50 Tensor 51 Tensor 52 Tensor 53 Tensor 53 Tensor 54 Tensor 54 Tensor 55 Tensor 56 Tensor 57 Tensor 57 Tensor 57 Tensor 58 Tensor 59 Tensor 59 Tensor 59 Tensor 59 Tensor 50 Tensor 50 Tensor 50 Tensor 50 Tensor 50 Tensor 51 Tensor 52 Tensor 53 Tensor 54 Tensor 54 Tensor 54 Tensor 55 Tensor 56 Tensor 57 Tensor 57 Tensor 58 Tensor 59 Tensor 59 Tensor 59 Tensor 50 Tensor	32.00K
Tensor29 Tensor30 Tensor30 Tensor31 Tensor32 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor38 Tensor39 Tensor39 Tensor40 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor48 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor49 Tensor49 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor5	32.00K
Tensor30 Tensor31 Tensor32 Tensor33 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor38 Tensor37 Tensor38 Tensor39 Tensor39 Tensor40 Tensor41 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor59 Ten	0.00B
Tensor31 Tensor32 Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor38 Tensor39 Tensor40 Tensor41 Tensor41 Tensor42 Tensor43 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor47 Tensor48 Tensor49 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor56 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor57 Tensor58 Tensor59 Ten	32.00K
Tensor32 Tensor33 Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor38 Tensor39 Tensor39 Tensor40 Tensor41 Tensor42 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor43 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor50 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor55 Tensor56 Tensor57 Tensor59 Ten	32.00K
Tensor33 Tensor34 Tensor35 Tensor35 Tensor36 Tensor37 Tensor38 Tensor38 Tensor39 Tensor39 Tensor40 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor49 Tensor49 Tensor49 Tensor40 Tensor40 Tensor41 Tensor41 Tensor42 Tensor43 Tensor45 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor49 Tensor50 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor53 Tensor53 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor57 Tensor58 Tensor59 Ten	0.00B
Tensor34 (1, 2, 4096) Tensor35 (1, 32, 2, 128) Tensor36 (1, 32, 2, 128) Tensor37 (1, 2, 4096) Tensor38 (1, 32, 2, 128) Tensor39 (1, 32, 2, 128) Tensor40 (1, 2, 4096) Tensor41 (1, 32, 2, 128) Tensor42 (1, 32, 2, 128) Tensor43 (1, 2, 4096) Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 32, 2, 128) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor35 Tensor36 Tensor37 Tensor37 Tensor38 Tensor39 Tensor40 Tensor41 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor48 Tensor49 Tensor50 Tensor50 Tensor51 Tensor51 Tensor52 Tensor53 Tensor53 Tensor54 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor56 Tensor57 Tensor57 Tensor57 Tensor57 Tensor58 Tensor59 Ten	32.00K
Tensor36 (1, 32, 2, 128) Tensor37 (1, 2, 4096) Tensor38 (1, 32, 2, 128) Tensor39 (1, 32, 2, 128) Tensor40 (1, 2, 4096) Tensor41 (1, 32, 2, 128) Tensor42 (1, 32, 2, 128) Tensor43 (1, 2, 4096) Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 32, 2, 128) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	0.00B
Tensor37 Tensor38 Tensor39 Tensor40 Tensor41 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor50 Tensor50 Tensor50 Tensor50 Tensor53 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor57 Tensor57 Tensor58 Tensor59 Ten	32.00K
Tensor38 (1, 32, 2, 128) Tensor39 (1, 32, 2, 128) Tensor40 (1, 2, 4096) Tensor41 (1, 32, 2, 128) Tensor42 (1, 32, 2, 128) Tensor43 (1, 2, 4096) Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 2, 4096) Tensor47 (1, 32, 2, 128) Tensor49 (1, 2, 4096) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor39 Tensor40 Tensor41 Tensor42 Tensor42 Tensor43 Tensor44 Tensor45 Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor50 Tensor50 Tensor50 Tensor51 Tensor52 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor54 Tensor55 Tensor55 Tensor55 Tensor55 Tensor56 Tensor57 Tensor57 Tensor58 Tensor59 Ten	0.00B
Tensor40 (1, 2, 4096) Tensor41 (1, 32, 2, 128) Tensor42 (1, 32, 2, 128) Tensor43 (1, 2, 4096) Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 2, 4096) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor41 (1, 32, 2, 128) Tensor42 (1, 32, 2, 128) Tensor43 (1, 2, 4096) Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 2, 4096) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor42 (1, 32, 2, 128) Tensor43 (1, 2, 4096) Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 2, 4096) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	0.00B
Tensor43 (1, 2, 4096) Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 2, 4096) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor50 (1, 2, 4096) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor44 (1, 32, 2, 128) Tensor45 (1, 32, 2, 128) Tensor46 (1, 2, 4096) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	
Tensor45 Tensor46 Tensor47 Tensor47 Tensor48 Tensor49 Tensor50 Tensor50 Tensor51 Tensor52 Tensor53 Tensor54 Ten	32.00K
Tensor46 (1, 2, 4096) Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor49 (1, 2, 4096) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	0.00B
Tensor47 (1, 32, 2, 128) Tensor48 (1, 32, 2, 128) Tensor49 (1, 2, 4096) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor48 (1, 32, 2, 128) Tensor49 (1, 2, 4096) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor49 (1, 2, 4096) Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	0.00B
Tensor50 (1, 32, 2, 128) Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor51 (1, 32, 2, 128) Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor52 (1, 2, 4096) Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	0.00B
Tensor53 (1, 32, 2, 128) Tensor54 (1, 32, 2, 128)	32.00K
Tensor54 (1, 32, 2, 128)	32.00K
	0.00B
TEF (4 0 4000)	32.00K
Tensor55 (1, 2, 4096) Tensor56 (1, 32, 2, 128)	32.00K 0.00B

```
Tensor57
                                               (1, 32, 2, 128)
                                                                   32.00K
Tensor58
                                                  (1, 2, 4096)
                                                                   32.00K
Tensor59
                                               (1, 32, 2, 128)
                                                                    0.00B
Tensor60
                                               (1, 32, 2, 128)
                                                                   32.00K
Tensor61
                                                   (1, 2, 4096)
                                                                   32.00K
                                               (1, 32, 2, 128)
Tensor62
                                                                    0.00B
Tensor63
                                               (1, 32, 2, 128)
                                                                   32.00K
                                                  (1, 2, 4096)
                                                                   32.00K
Tensor64
                                               (1, 32, 2, 128)
Tensor65
                                                                    0.00B
Tensor66
                                               (1, 32, 2, 128)
                                                                   32.00K
Tensor67
                                                  (1, 2, 4096)
                                                                   32.00K
                                               (1, 32, 2, 128)
Tensor68
                                                                    0.00B
                                               (1, 32, 2, 128)
Tensor69
                                                                   32.00K
Tensor70
                                                   (1, 2, 4096)
                                                                   32.00K
Tensor71
                                               (1, 32, 2, 128)
                                                                    0.00B
Tensor72
                                               (1, 32, 2, 128)
                                                                   32.00K
Tensor73
                                                  (1, 2, 4096)
                                                                   32.00K
                                               (1, 32, 2, 128)
Tensor74
                                                                    0.00B
Tensor75
                                               (1, 32, 2, 128)
                                                                   32.00K
Tensor76
                                                   (1, 2, 4096)
                                                                   32.00K
                                               (1, 32, 2, 128)
Tensor77
                                                                    0.00B
Tensor78
                                               (1, 32, 2, 128)
                                                                   32.00K
Tensor79
                                                  (1, 2, 4096)
                                                                   32.00K
                                               (1, 32, 2, 128)
                                                                    0.00B
Tensor80
Tensor81
                                               (1, 32, 2, 128)
                                                                   32.00K
Tensor82
                                                  (1, 2, 4096)
                                                                   32.00K
                                               (1, 32, 2, 128)
Tensor83
                                                                    0.00B
Tensor84
                                               (1, 32, 2, 128)
                                                                   32.00K
                                                   (1, 2, 4096)
Tensor85
                                                                   32.00K
Tensor86
                                               (1, 32, 2, 128)
                                                                    0.00B
                                               (1, 32, 2, 128)
Tensor87
                                                                   32.00K
Tensor88
                                                  (1, 2, 4096)
                                                                   32.00K
                                                                    0.00B
Tensor89
                                               (1, 32, 2, 128)
Tensor90
                                               (1, 32, 2, 128)
                                                                   32.00K
                                                   (1, 2, 4096)
Tensor91
                                                                   32.00K
Tensor92
                                               (1, 32, 2, 128)
                                                                    0.00B
Tensor93
                                               (1, 32, 2, 128)
                                                                   32.00K
                                                  (1, 2, 4096)
Tensor94
                                                                   32.00K
Tensor95
                                               (1, 32, 2, 128)
                                                                    0.00B
                                               (1, 32, 2, 128)
Tensor96
                                                                   32.00K
Tensor97
                                                 (1, 2, 32000)
                                                                  250.00K
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
                                                              (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
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
                                                                   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
                                                                  512.00B
Tensor117
                                                         (64,)
Tensor118
                                                         (64,)
                                                                  512.00B
Tensor119
                                                         (64,)
                                                                  512.00B
Tensor120
                                                         (64,)
                                                                  512.00B
Tensor121
                                                         (64,)
                                                                  512.00B
Tensor122
                                                         (64,)
                                                                  512.00B
Tensor123
                                                                  512.00B
                                                         (64,)
Tensor124
                                                         (64,)
                                                                  512.00B
Tensor125
                                                         (64,)
                                                                  512.00B
Tensor126
                                                         (64,)
                                                                  512.00B
Tensor127
                                                                  512.00B
                                                         (64,)
Tensor128
                                                                  512.00B
                                                         (64,)
Tensor129
                                                                  512.00B
                                                         (64,)
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
                                                  (4096, 4096)
model.layers.0.self_attn.v_proj.weight
                                                                   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
                                                  (4096, 4096)
model.layers.1.self_attn.v_proj.weight
                                                                   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
                                                 (4096, 11008)
model.layers.1.mlp.down proj.weight
                                                                  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
                                                  (4096, 4096)
model.layers.2.self_attn.v_proj.weight
                                                                   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
                                                 (11008, 4096)
model.layers.2.mlp.up proj.weight
                                                                  172.00M
model.layers.2.mlp.down_proj.weight
                                                 (4096, 11008)
                                                                  172.00M
                                                  (4096, 4096)
model.layers.3.self_attn.q_proj.weight
                                                                   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
                                                 (4096, 11008)
model.layers.3.mlp.down_proj.weight
                                                                  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
                                                 (11008, 4096)
model.layers.4.mlp.gate proj.weight
                                                                  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
                                                  (4096, 4096)
                                                                   64.00M
model.layers.5.self_attn.o_proj.weight
```

```
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
                                                 (11008, 4096)
model.layers.6.mlp.up_proj.weight
                                                                 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
                                                 (4096, 4096)
model.layers.7.self_attn.k_proj.weight
                                                                  64.00M
                                                 (4096, 4096)
model.layers.7.self_attn.v_proj.weight
                                                                  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
                                                 (11008, 4096)
model.layers.7.mlp.up_proj.weight
                                                                 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
                                                 (4096, 4096)
model.layers.8.self_attn.o_proj.weight
                                                                  64.00M
model.layers.8.mlp.gate_proj.weight
                                                 (11008, 4096)
                                                                 172.00M
model.layers.8.mlp.up_proj.weight
                                                 (11008, 4096)
                                                                 172.00M
                                                 (4096, 11008)
                                                                 172.00M
model.layers.8.mlp.down_proj.weight
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
                                                 (4096, 4096)
model.layers.10.self_attn.v_proj.weight
                                                                  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
                                                 (4096, 11008)
model.layers.10.mlp.down_proj.weight
                                                                 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
                                                 (4096, 4096)
model.layers.11.self_attn.o_proj.weight
                                                                  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
                                                  (4096, 4096)
model.layers.13.self_attn.k_proj.weight
                                                                  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
                                                 (11008, 4096)
model.layers.13.mlp.gate_proj.weight
                                                                 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
                                                 (4096, 11008)
model.layers.15.mlp.down_proj.weight
                                                                 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
                                                 (4096, 4096)
model.layers.16.self_attn.v_proj.weight
                                                                  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
                                                 (4096, 4096)
model.layers.17.self_attn.o_proj.weight
                                                                  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
                                                 (4096, 4096)
model.layers.18.self_attn.o_proj.weight
                                                                  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
                                                 (4096, 4096)
model.layers.19.self_attn.o_proj.weight
                                                                  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
                                                 (11008, 4096)
model.layers.20.mlp.gate_proj.weight
                                                                 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
                                                 (11008, 4096)
model.layers.21.mlp.gate proj.weight
                                                                 172.00M
                                                 (11008, 4096)
model.layers.21.mlp.up proj.weight
                                                                 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
                                                 (4096, 11008)
model.layers.22.mlp.down_proj.weight
                                                                 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
                                                 (4096, 11008)
model.layers.23.mlp.down_proj.weight
                                                                 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
                                                 (4096, 4096)
model.layers.24.self_attn.v_proj.weight
                                                                  64.00M
                                                 (4096, 4096)
model.layers.24.self_attn.o_proj.weight
                                                                  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
                                                 (11008, 4096)
model.layers.28.mlp.up_proj.weight
                                                                 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
                                                  (4096, 4096)
model.layers.30.self_attn.v_proj.weight
                                                                  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
                                                 (11008, 4096)
model.layers.30.mlp.up_proj.weight
                                                                 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)
                                        (4096, 11008)
model.layers.31.mlp.down_proj.weight
                                                       172.00M
______
Total Tensors: 6739268162
                          Used Memory: 25.10G
______
Logits shape: torch.Size([1, 2, 32000])
Logits: tensor([[[-12.9832, -7.4134, -0.4327, ..., -6.8297, -8.0880, -7.586
       [-10.4322, -6.7053, -0.7051, \ldots, -6.9910, -6.6368, -8.2753]]])
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 []: