In [1]:
```
from transformers import AutoModelForCausalLM, AutoTokenizer
import torch
from torch.amp import autocast
from pytorch_memlab import MemReporter
import gc
```

In [2]:
```
model_directory = r"Z:/llmfile/Llama-2-7B-hf/models--meta-llama--Llama-2-7b-hf/snapshots/01c7f73d771dfac7d29232

model = AutoModelForCausalLM.from_pretrained(model_directory)
tokenizer = AutoTokenizer.from_pretrained(model_directory)

# 初始输入
input_text = "Hello"
input_ids = tokenizer(input_text, return_tensors="pt").input_ids

reporter = MemReporter(model)

# 前向传播
print("==ForwardTrainingMemoryUsage==")
model.train()
with autocast("cuda"): # 启用混合精度
    outputs = model(input_ids)
    loss = outputs.loss
    reporter.report()  # 前向传播后的内存使用情况
```

```
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
==ForwardTrainingMemoryUsage==
Element type                                            Size  Used MEM
-------------------------------------------------------------------------------
Storage on cpu
Tensor0                                        (1, 2, 11008)    86.00K
Tensor1                                        (1, 2, 11008)    86.00K
Tensor2                                        (1, 2, 11008)    86.00K
Tensor3                                        (1, 2, 11008)    86.00K
Tensor4                                         (1, 2, 4096)    32.00K
Tensor5                                            (1, 2, 1)   512.00B
Tensor6                                         (1, 2, 4096)    32.00K
Tensor7                                         (1, 2, 4096)    32.00K
Tensor8                                         (1, 2, 4096)    32.00K
Tensor9                                      (1, 32, 2, 128)     0.00B
Tensor10                                      (1, 1, 2, 128)     1.00K
Tensor11                                      (1, 1, 2, 128)     1.00K
Tensor12                                     (1, 32, 2, 128)    32.00K
Tensor13                                     (1, 32, 2, 128)    32.00K
Tensor14                                     (1, 32, 2, 128)    32.00K
Tensor15                                        (1, 2, 4096)    32.00K
Tensor16                                           (1, 2, 1)   512.00B
Tensor17                                        (1, 2, 4096)    32.00K
Tensor18                                        (1, 2, 4096)    32.00K
Tensor19                                       (1, 2, 11008)    86.00K
Tensor20                                       (1, 2, 11008)    86.00K
Tensor21                                       (1, 2, 11008)    86.00K
Tensor22                                       (1, 2, 11008)    86.00K
Tensor23                                        (1, 2, 4096)    32.00K
Tensor24                                           (1, 2, 1)   512.00B
Tensor25                                        (1, 2, 4096)    32.00K
Tensor26                                        (1, 2, 4096)    32.00K
Tensor27                                        (1, 2, 4096)    32.00K
Tensor28                                     (1, 32, 2, 128)     0.00B
Tensor29                                      (1, 1, 2, 128)     0.00B
Tensor30                                      (1, 1, 2, 128)     0.00B
Tensor31                                     (1, 32, 2, 128)    32.00K
Tensor32                                     (1, 32, 2, 128)    32.00K
Tensor33                                     (1, 32, 2, 128)    32.00K
Tensor34                                        (1, 2, 4096)    32.00K
Tensor35                                           (1, 2, 1)   512.00B
Tensor36                                        (1, 2, 4096)    32.00K
Tensor37                                        (1, 2, 4096)    32.00K
Tensor38                                       (1, 2, 11008)    86.00K
Tensor39                                       (1, 2, 11008)    86.00K
Tensor40                                       (1, 2, 11008)    86.00K
Tensor41                                       (1, 2, 11008)    86.00K
Tensor42                                        (1, 2, 4096)    32.00K
Tensor43                                           (1, 2, 1)   512.00B
Tensor44                                        (1, 2, 4096)    32.00K
Tensor45                                        (1, 2, 4096)    32.00K
Tensor46                                        (1, 2, 4096)    32.00K
Tensor47                                     (1, 32, 2, 128)     0.00B
Tensor48                                      (1, 1, 2, 128)     0.00B
Tensor49                                      (1, 1, 2, 128)     0.00B
Tensor50                                     (1, 32, 2, 128)    32.00K
Tensor51                                     (1, 32, 2, 128)    32.00K
```

```
Tensor52                                     (1, 32, 2, 128)    32.00K
Tensor53                                        (1, 2, 4096)    32.00K
Tensor54                                           (1, 2, 1)   512.00B
Tensor55                                        (1, 2, 4096)    32.00K
Tensor56                                        (1, 2, 4096)    32.00K
Tensor57                                       (1, 2, 11008)    86.00K
Tensor58                                       (1, 2, 11008)    86.00K
Tensor59                                       (1, 2, 11008)    86.00K
Tensor60                                       (1, 2, 11008)    86.00K
Tensor61                                        (1, 2, 4096)    32.00K
Tensor62                                           (1, 2, 1)   512.00B
Tensor63                                        (1, 2, 4096)    32.00K
Tensor64                                        (1, 2, 4096)    32.00K
Tensor65                                        (1, 2, 4096)    32.00K
Tensor66                                     (1, 32, 2, 128)     0.00B
Tensor67                                      (1, 1, 2, 128)     0.00B
Tensor68                                      (1, 1, 2, 128)     0.00B
Tensor69                                     (1, 32, 2, 128)    32.00K
Tensor70                                     (1, 32, 2, 128)    32.00K
Tensor71                                     (1, 32, 2, 128)    32.00K
Tensor72                                        (1, 2, 4096)    32.00K
Tensor73                                           (1, 2, 1)   512.00B
Tensor74                                        (1, 2, 4096)    32.00K
Tensor75                                        (1, 2, 4096)    32.00K
Tensor76                                       (1, 2, 11008)    86.00K
Tensor77                                       (1, 2, 11008)    86.00K
Tensor78                                       (1, 2, 11008)    86.00K
Tensor79                                       (1, 2, 11008)    86.00K
Tensor80                                        (1, 2, 4096)    32.00K
Tensor81                                           (1, 2, 1)   512.00B
Tensor82                                        (1, 2, 4096)    32.00K
Tensor83                                        (1, 2, 4096)    32.00K
Tensor84                                        (1, 2, 4096)    32.00K
Tensor85                                     (1, 32, 2, 128)     0.00B
Tensor86                                      (1, 1, 2, 128)     0.00B
Tensor87                                      (1, 1, 2, 128)     0.00B
Tensor88                                     (1, 32, 2, 128)    32.00K
Tensor89                                     (1, 32, 2, 128)    32.00K
Tensor90                                     (1, 32, 2, 128)    32.00K
Tensor91                                        (1, 2, 4096)    32.00K
Tensor92                                           (1, 2, 1)   512.00B
Tensor93                                        (1, 2, 4096)    32.00K
Tensor94                                        (1, 2, 4096)    32.00K
Tensor95                                       (1, 2, 11008)    86.00K
Tensor96                                       (1, 2, 11008)    86.00K
Tensor97                                       (1, 2, 11008)    86.00K
Tensor98                                       (1, 2, 11008)    86.00K
Tensor99                                        (1, 2, 4096)    32.00K
Tensor100                                          (1, 2, 1)   512.00B
Tensor101                                       (1, 2, 4096)    32.00K
Tensor102                                       (1, 2, 4096)    32.00K
Tensor103                                       (1, 2, 4096)    32.00K
Tensor104                                    (1, 32, 2, 128)     0.00B
Tensor105                                     (1, 1, 2, 128)     0.00B
Tensor106                                     (1, 1, 2, 128)     0.00B
Tensor107                                    (1, 32, 2, 128)    32.00K
Tensor108                                    (1, 32, 2, 128)    32.00K
Tensor109                                    (1, 32, 2, 128)    32.00K
Tensor110                                       (1, 2, 4096)    32.00K
Tensor111                                          (1, 2, 1)   512.00B
Tensor112                                       (1, 2, 4096)    32.00K
Tensor113                                       (1, 2, 4096)    32.00K
Tensor114                                      (1, 2, 11008)    86.00K
Tensor115                                      (1, 2, 11008)    86.00K
Tensor116                                      (1, 2, 11008)    86.00K
Tensor117                                      (1, 2, 11008)    86.00K
Tensor118                                       (1, 2, 4096)    32.00K
Tensor119                                          (1, 2, 1)   512.00B
Tensor120                                       (1, 2, 4096)    32.00K
Tensor121                                       (1, 2, 4096)    32.00K
Tensor122                                       (1, 2, 4096)    32.00K
Tensor123                                    (1, 32, 2, 128)     0.00B
Tensor124                                     (1, 1, 2, 128)     0.00B
Tensor125                                     (1, 1, 2, 128)     0.00B
Tensor126                                    (1, 32, 2, 128)    32.00K
Tensor127                                    (1, 32, 2, 128)    32.00K
Tensor128                                    (1, 32, 2, 128)    32.00K
Tensor129                                       (1, 2, 4096)    32.00K
Tensor130                                          (1, 2, 1)   512.00B
Tensor131                                       (1, 2, 4096)    32.00K
Tensor132                                       (1, 2, 4096)    32.00K
Tensor133                                      (1, 2, 11008)    86.00K
Tensor134                                      (1, 2, 11008)    86.00K
```

```
Tensor135                                      (1, 2, 11008)    86.00K
Tensor136                                      (1, 2, 11008)    86.00K
Tensor137                                       (1, 2, 4096)    32.00K
Tensor138                                          (1, 2, 1)   512.00B
Tensor139                                       (1, 2, 4096)    32.00K
Tensor140                                       (1, 2, 4096)    32.00K
Tensor141                                       (1, 2, 4096)    32.00K
Tensor142                                    (1, 32, 2, 128)     0.00B
Tensor143                                     (1, 1, 2, 128)     0.00B
Tensor144                                     (1, 1, 2, 128)     0.00B
Tensor145                                    (1, 32, 2, 128)    32.00K
Tensor146                                    (1, 32, 2, 128)    32.00K
Tensor147                                    (1, 32, 2, 128)    32.00K
Tensor148                                       (1, 2, 4096)    32.00K
Tensor149                                          (1, 2, 1)   512.00B
Tensor150                                       (1, 2, 4096)    32.00K
Tensor151                                       (1, 2, 4096)    32.00K
Tensor152                                      (1, 2, 11008)    86.00K
Tensor153                                      (1, 2, 11008)    86.00K
Tensor154                                      (1, 2, 11008)    86.00K
Tensor155                                      (1, 2, 11008)    86.00K
Tensor156                                       (1, 2, 4096)    32.00K
Tensor157                                          (1, 2, 1)   512.00B
Tensor158                                       (1, 2, 4096)    32.00K
Tensor159                                       (1, 2, 4096)    32.00K
Tensor160                                       (1, 2, 4096)    32.00K
Tensor161                                    (1, 32, 2, 128)     0.00B
Tensor162                                     (1, 1, 2, 128)     0.00B
Tensor163                                     (1, 1, 2, 128)     0.00B
Tensor164                                    (1, 32, 2, 128)    32.00K
Tensor165                                    (1, 32, 2, 128)    32.00K
Tensor166                                    (1, 32, 2, 128)    32.00K
Tensor167                                       (1, 2, 4096)    32.00K
Tensor168                                          (1, 2, 1)   512.00B
Tensor169                                       (1, 2, 4096)    32.00K
Tensor170                                       (1, 2, 4096)    32.00K
Tensor171                                      (1, 2, 11008)    86.00K
Tensor172                                      (1, 2, 11008)    86.00K
Tensor173                                      (1, 2, 11008)    86.00K
Tensor174                                      (1, 2, 11008)    86.00K
Tensor175                                       (1, 2, 4096)    32.00K
Tensor176                                          (1, 2, 1)   512.00B
Tensor177                                       (1, 2, 4096)    32.00K
Tensor178                                       (1, 2, 4096)    32.00K
Tensor179                                       (1, 2, 4096)    32.00K
Tensor180                                    (1, 32, 2, 128)     0.00B
Tensor181                                     (1, 1, 2, 128)     0.00B
Tensor182                                     (1, 1, 2, 128)     0.00B
Tensor183                                    (1, 32, 2, 128)    32.00K
Tensor184                                    (1, 32, 2, 128)    32.00K
Tensor185                                    (1, 32, 2, 128)    32.00K
Tensor186                                       (1, 2, 4096)    32.00K
Tensor187                                          (1, 2, 1)   512.00B
Tensor188                                       (1, 2, 4096)    32.00K
Tensor189                                       (1, 2, 4096)    32.00K
Tensor190                                      (1, 2, 11008)    86.00K
Tensor191                                      (1, 2, 11008)    86.00K
Tensor192                                      (1, 2, 11008)    86.00K
Tensor193                                      (1, 2, 11008)    86.00K
Tensor194                                       (1, 2, 4096)    32.00K
Tensor195                                          (1, 2, 1)   512.00B
Tensor196                                       (1, 2, 4096)    32.00K
Tensor197                                       (1, 2, 4096)    32.00K
Tensor198                                       (1, 2, 4096)    32.00K
Tensor199                                    (1, 32, 2, 128)     0.00B
Tensor200                                     (1, 1, 2, 128)     0.00B
Tensor201                                     (1, 1, 2, 128)     0.00B
Tensor202                                    (1, 32, 2, 128)    32.00K
Tensor203                                    (1, 32, 2, 128)    32.00K
Tensor204                                    (1, 32, 2, 128)    32.00K
Tensor205                                       (1, 2, 4096)    32.00K
Tensor206                                          (1, 2, 1)   512.00B
Tensor207                                       (1, 2, 4096)    32.00K
Tensor208                                       (1, 2, 4096)    32.00K
Tensor209                                      (1, 2, 11008)    86.00K
Tensor210                                      (1, 2, 11008)    86.00K
Tensor211                                      (1, 2, 11008)    86.00K
Tensor212                                      (1, 2, 11008)    86.00K
Tensor213                                       (1, 2, 4096)    32.00K
Tensor214                                          (1, 2, 1)   512.00B
Tensor215                                       (1, 2, 4096)    32.00K
Tensor216                                       (1, 2, 4096)    32.00K
Tensor217                                       (1, 2, 4096)    32.00K
```

```
Tensor218                                    (1, 32, 2, 128)     0.00B
Tensor219                                     (1, 1, 2, 128)     0.00B
Tensor220                                     (1, 1, 2, 128)     0.00B
Tensor221                                    (1, 32, 2, 128)    32.00K
Tensor222                                    (1, 32, 2, 128)    32.00K
Tensor223                                    (1, 32, 2, 128)    32.00K
Tensor224                                       (1, 2, 4096)    32.00K
Tensor225                                          (1, 2, 1)   512.00B
Tensor226                                       (1, 2, 4096)    32.00K
Tensor227                                       (1, 2, 4096)    32.00K
Tensor228                                      (1, 2, 11008)    86.00K
Tensor229                                      (1, 2, 11008)    86.00K
Tensor230                                      (1, 2, 11008)    86.00K
Tensor231                                      (1, 2, 11008)    86.00K
Tensor232                                       (1, 2, 4096)    32.00K
Tensor233                                          (1, 2, 1)   512.00B
Tensor234                                       (1, 2, 4096)    32.00K
Tensor235                                       (1, 2, 4096)    32.00K
Tensor236                                       (1, 2, 4096)    32.00K
Tensor237                                    (1, 32, 2, 128)     0.00B
Tensor238                                     (1, 1, 2, 128)     0.00B
Tensor239                                     (1, 1, 2, 128)     0.00B
Tensor240                                    (1, 32, 2, 128)    32.00K
Tensor241                                    (1, 32, 2, 128)    32.00K
Tensor242                                    (1, 32, 2, 128)    32.00K
Tensor243                                       (1, 2, 4096)    32.00K
Tensor244                                          (1, 2, 1)   512.00B
Tensor245                                       (1, 2, 4096)    32.00K
Tensor246                                       (1, 2, 4096)    32.00K
Tensor247                                      (1, 2, 11008)    86.00K
Tensor248                                      (1, 2, 11008)    86.00K
Tensor249                                      (1, 2, 11008)    86.00K
Tensor250                                      (1, 2, 11008)    86.00K
Tensor251                                       (1, 2, 4096)    32.00K
Tensor252                                          (1, 2, 1)   512.00B
Tensor253                                       (1, 2, 4096)    32.00K
Tensor254                                       (1, 2, 4096)    32.00K
Tensor255                                       (1, 2, 4096)    32.00K
Tensor256                                    (1, 32, 2, 128)     0.00B
Tensor257                                     (1, 1, 2, 128)     0.00B
Tensor258                                     (1, 1, 2, 128)     0.00B
Tensor259                                    (1, 32, 2, 128)    32.00K
Tensor260                                    (1, 32, 2, 128)    32.00K
Tensor261                                    (1, 32, 2, 128)    32.00K
Tensor262                                       (1, 2, 4096)    32.00K
Tensor263                                          (1, 2, 1)   512.00B
Tensor264                                       (1, 2, 4096)    32.00K
Tensor265                                       (1, 2, 4096)    32.00K
Tensor266                                      (1, 2, 11008)    86.00K
Tensor267                                      (1, 2, 11008)    86.00K
Tensor268                                      (1, 2, 11008)    86.00K
Tensor269                                      (1, 2, 11008)    86.00K
Tensor270                                       (1, 2, 4096)    32.00K
Tensor271                                          (1, 2, 1)   512.00B
Tensor272                                       (1, 2, 4096)    32.00K
Tensor273                                       (1, 2, 4096)    32.00K
Tensor274                                       (1, 2, 4096)    32.00K
Tensor275                                    (1, 32, 2, 128)     0.00B
Tensor276                                     (1, 1, 2, 128)     0.00B
Tensor277                                     (1, 1, 2, 128)     0.00B
Tensor278                                    (1, 32, 2, 128)    32.00K
Tensor279                                    (1, 32, 2, 128)    32.00K
Tensor280                                    (1, 32, 2, 128)    32.00K
Tensor281                                       (1, 2, 4096)    32.00K
Tensor282                                          (1, 2, 1)   512.00B
Tensor283                                       (1, 2, 4096)    32.00K
Tensor284                                       (1, 2, 4096)    32.00K
Tensor285                                      (1, 2, 11008)    86.00K
Tensor286                                      (1, 2, 11008)    86.00K
Tensor287                                      (1, 2, 11008)    86.00K
Tensor288                                      (1, 2, 11008)    86.00K
Tensor289                                       (1, 2, 4096)    32.00K
Tensor290                                          (1, 2, 1)   512.00B
Tensor291                                       (1, 2, 4096)    32.00K
Tensor292                                       (1, 2, 4096)    32.00K
Tensor293                                      (1, 2, 32000)   250.00K
Tensor294                                             (1, 2)   512.00B
Tensor295                                       (1, 2, 4096)    32.00K
Tensor296                                          (1, 2, 1)   512.00B
Tensor297                                       (1, 2, 4096)    32.00K
Tensor298                                       (1, 2, 4096)    32.00K
Tensor299                                       (1, 2, 4096)    32.00K
Tensor300                                    (1, 32, 2, 128)     0.00B
```

```
Tensor301                                     (1, 1, 2, 128)     0.00B
Tensor302                                     (1, 1, 2, 128)     0.00B
Tensor303                                    (1, 32, 2, 128)    32.00K
Tensor304                                    (1, 32, 2, 128)    32.00K
Tensor305                                    (1, 32, 2, 128)    32.00K
Tensor306                                       (1, 2, 4096)    32.00K
Tensor307                                          (1, 2, 1)   512.00B
Tensor308                                       (1, 2, 4096)    32.00K
Tensor309                                       (1, 2, 4096)    32.00K
Tensor310                                      (1, 2, 11008)    86.00K
Tensor311                                      (1, 2, 11008)    86.00K
Tensor312                                      (1, 2, 11008)    86.00K
Tensor313                                      (1, 2, 11008)    86.00K
Tensor314                                       (1, 2, 4096)    32.00K
Tensor315                                          (1, 2, 1)   512.00B
Tensor316                                       (1, 2, 4096)    32.00K
Tensor317                                       (1, 2, 4096)    32.00K
Tensor318                                       (1, 2, 4096)    32.00K
Tensor319                                    (1, 32, 2, 128)     0.00B
Tensor320                                     (1, 1, 2, 128)     0.00B
Tensor321                                     (1, 1, 2, 128)     0.00B
Tensor322                                    (1, 32, 2, 128)    32.00K
Tensor323                                    (1, 32, 2, 128)    32.00K
Tensor324                                    (1, 32, 2, 128)    32.00K
Tensor325                                       (1, 2, 4096)    32.00K
Tensor326                                          (1, 2, 1)   512.00B
Tensor327                                       (1, 2, 4096)    32.00K
Tensor328                                       (1, 2, 4096)    32.00K
Tensor329                                      (1, 2, 11008)    86.00K
Tensor330                                      (1, 2, 11008)    86.00K
Tensor331                                      (1, 2, 11008)    86.00K
Tensor332                                      (1, 2, 11008)    86.00K
Tensor333                                       (1, 2, 4096)    32.00K
Tensor334                                          (1, 2, 1)   512.00B
Tensor335                                       (1, 2, 4096)    32.00K
Tensor336                                       (1, 2, 4096)    32.00K
Tensor337                                       (1, 2, 4096)    32.00K
Tensor338                                    (1, 32, 2, 128)     0.00B
Tensor339                                     (1, 1, 2, 128)     0.00B
Tensor340                                     (1, 1, 2, 128)     0.00B
Tensor341                                    (1, 32, 2, 128)    32.00K
Tensor342                                    (1, 32, 2, 128)    32.00K
Tensor343                                    (1, 32, 2, 128)    32.00K
Tensor344                                       (1, 2, 4096)    32.00K
Tensor345                                          (1, 2, 1)   512.00B
Tensor346                                       (1, 2, 4096)    32.00K
Tensor347                                       (1, 2, 4096)    32.00K
Tensor348                                      (1, 2, 11008)    86.00K
Tensor349                                      (1, 2, 11008)    86.00K
Tensor350                                      (1, 2, 11008)    86.00K
Tensor351                                      (1, 2, 11008)    86.00K
Tensor352                                       (1, 2, 4096)    32.00K
Tensor353                                          (1, 2, 1)   512.00B
Tensor354                                       (1, 2, 4096)    32.00K
Tensor355                                       (1, 2, 4096)    32.00K
Tensor356                                       (1, 2, 4096)    32.00K
Tensor357                                    (1, 32, 2, 128)     0.00B
Tensor358                                     (1, 1, 2, 128)     0.00B
Tensor359                                     (1, 1, 2, 128)     0.00B
Tensor360                                    (1, 32, 2, 128)    32.00K
Tensor361                                    (1, 32, 2, 128)    32.00K
Tensor362                                    (1, 32, 2, 128)    32.00K
Tensor363                                       (1, 2, 4096)    32.00K
Tensor364                                          (1, 2, 1)   512.00B
Tensor365                                       (1, 2, 4096)    32.00K
Tensor366                                       (1, 2, 4096)    32.00K
Tensor367                                      (1, 2, 11008)    86.00K
Tensor368                                      (1, 2, 11008)    86.00K
Tensor369                                      (1, 2, 11008)    86.00K
Tensor370                                      (1, 2, 11008)    86.00K
Tensor371                                       (1, 2, 4096)    32.00K
Tensor372                                          (1, 2, 1)   512.00B
Tensor373                                       (1, 2, 4096)    32.00K
Tensor374                                       (1, 2, 4096)    32.00K
Tensor375                                       (1, 2, 4096)    32.00K
Tensor376                                    (1, 32, 2, 128)     0.00B
Tensor377                                     (1, 1, 2, 128)     0.00B
Tensor378                                     (1, 1, 2, 128)     0.00B
Tensor379                                    (1, 32, 2, 128)    32.00K
Tensor380                                    (1, 32, 2, 128)    32.00K
Tensor381                                    (1, 32, 2, 128)    32.00K
Tensor382                                       (1, 2, 4096)    32.00K
Tensor383                                          (1, 2, 1)   512.00B
```

```
Tensor384                                       (1, 2, 4096)    32.00K
Tensor385                                       (1, 2, 4096)    32.00K
Tensor386                                      (1, 2, 11008)    86.00K
Tensor387                                      (1, 2, 11008)    86.00K
Tensor388                                      (1, 2, 11008)    86.00K
Tensor389                                      (1, 2, 11008)    86.00K
Tensor390                                       (1, 2, 4096)    32.00K
Tensor391                                          (1, 2, 1)   512.00B
Tensor392                                       (1, 2, 4096)    32.00K
Tensor393                                       (1, 2, 4096)    32.00K
Tensor394                                       (1, 2, 4096)    32.00K
Tensor395                                    (1, 32, 2, 128)     0.00B
Tensor396                                     (1, 1, 2, 128)     0.00B
Tensor397                                     (1, 1, 2, 128)     0.00B
Tensor398                                    (1, 32, 2, 128)    32.00K
Tensor399                                    (1, 32, 2, 128)    32.00K
Tensor400                                    (1, 32, 2, 128)    32.00K
Tensor401                                       (1, 2, 4096)    32.00K
Tensor402                                          (1, 2, 1)   512.00B
Tensor403                                       (1, 2, 4096)    32.00K
Tensor404                                       (1, 2, 4096)    32.00K
Tensor405                                      (1, 2, 11008)    86.00K
Tensor406                                      (1, 2, 11008)    86.00K
Tensor407                                      (1, 2, 11008)    86.00K
Tensor408                                      (1, 2, 11008)    86.00K
Tensor409                                       (1, 2, 4096)    32.00K
Tensor410                                          (1, 2, 1)   512.00B
Tensor411                                       (1, 2, 4096)    32.00K
Tensor412                                       (1, 2, 4096)    32.00K
Tensor413                                       (1, 2, 4096)    32.00K
Tensor414                                    (1, 32, 2, 128)     0.00B
Tensor415                                     (1, 1, 2, 128)     0.00B
Tensor416                                     (1, 1, 2, 128)     0.00B
Tensor417                                    (1, 32, 2, 128)    32.00K
Tensor418                                    (1, 32, 2, 128)    32.00K
Tensor419                                    (1, 32, 2, 128)    32.00K
Tensor420                                       (1, 2, 4096)    32.00K
Tensor421                                          (1, 2, 1)   512.00B
Tensor422                                       (1, 2, 4096)    32.00K
Tensor423                                       (1, 2, 4096)    32.00K
Tensor424                                      (1, 2, 11008)    86.00K
Tensor425                                      (1, 2, 11008)    86.00K
Tensor426                                      (1, 2, 11008)    86.00K
Tensor427                                      (1, 2, 11008)    86.00K
Tensor428                                       (1, 2, 4096)    32.00K
Tensor429                                          (1, 2, 1)   512.00B
Tensor430                                       (1, 2, 4096)    32.00K
Tensor431                                       (1, 2, 4096)    32.00K
Tensor432                                       (1, 2, 4096)    32.00K
Tensor433                                    (1, 32, 2, 128)     0.00B
Tensor434                                     (1, 1, 2, 128)     0.00B
Tensor435                                     (1, 1, 2, 128)     0.00B
Tensor436                                    (1, 32, 2, 128)    32.00K
Tensor437                                    (1, 32, 2, 128)    32.00K
Tensor438                                    (1, 32, 2, 128)    32.00K
Tensor439                                       (1, 2, 4096)    32.00K
Tensor440                                          (1, 2, 1)   512.00B
Tensor441                                       (1, 2, 4096)    32.00K
Tensor442                                       (1, 2, 4096)    32.00K
Tensor443                                      (1, 2, 11008)    86.00K
Tensor444                                      (1, 2, 11008)    86.00K
Tensor445                                      (1, 2, 11008)    86.00K
Tensor446                                      (1, 2, 11008)    86.00K
Tensor447                                       (1, 2, 4096)    32.00K
Tensor448                                          (1, 2, 1)   512.00B
Tensor449                                       (1, 2, 4096)    32.00K
Tensor450                                       (1, 2, 4096)    32.00K
Tensor451                                       (1, 2, 4096)    32.00K
Tensor452                                    (1, 32, 2, 128)     0.00B
Tensor453                                     (1, 1, 2, 128)     0.00B
Tensor454                                     (1, 1, 2, 128)     0.00B
Tensor455                                    (1, 32, 2, 128)    32.00K
Tensor456                                    (1, 32, 2, 128)    32.00K
Tensor457                                    (1, 32, 2, 128)    32.00K
Tensor458                                       (1, 2, 4096)    32.00K
Tensor459                                          (1, 2, 1)   512.00B
Tensor460                                       (1, 2, 4096)    32.00K
Tensor461                                       (1, 2, 4096)    32.00K
Tensor462                                      (1, 2, 11008)    86.00K
Tensor463                                      (1, 2, 11008)    86.00K
Tensor464                                      (1, 2, 11008)    86.00K
Tensor465                                      (1, 2, 11008)    86.00K
Tensor466                                       (1, 2, 4096)    32.00K
```

```
Tensor467                                          (1, 2, 1)   512.00B
Tensor468                                       (1, 2, 4096)    32.00K
Tensor469                                       (1, 2, 4096)    32.00K
Tensor470                                       (1, 2, 4096)    32.00K
Tensor471                                    (1, 32, 2, 128)     0.00B
Tensor472                                     (1, 1, 2, 128)     0.00B
Tensor473                                     (1, 1, 2, 128)     0.00B
Tensor474                                    (1, 32, 2, 128)    32.00K
Tensor475                                    (1, 32, 2, 128)    32.00K
Tensor476                                    (1, 32, 2, 128)    32.00K
Tensor477                                       (1, 2, 4096)    32.00K
Tensor478                                          (1, 2, 1)   512.00B
Tensor479                                       (1, 2, 4096)    32.00K
Tensor480                                       (1, 2, 4096)    32.00K
Tensor481                                      (1, 2, 11008)    86.00K
Tensor482                                      (1, 2, 11008)    86.00K
Tensor483                                      (1, 2, 11008)    86.00K
Tensor484                                      (1, 2, 11008)    86.00K
Tensor485                                       (1, 2, 4096)    32.00K
Tensor486                                          (1, 2, 1)   512.00B
Tensor487                                       (1, 2, 4096)    32.00K
Tensor488                                       (1, 2, 4096)    32.00K
Tensor489                                       (1, 2, 4096)    32.00K
Tensor490                                    (1, 32, 2, 128)     0.00B
Tensor491                                     (1, 1, 2, 128)     0.00B
Tensor492                                     (1, 1, 2, 128)     0.00B
Tensor493                                    (1, 32, 2, 128)    32.00K
Tensor494                                    (1, 32, 2, 128)    32.00K
Tensor495                                    (1, 32, 2, 128)    32.00K
Tensor496                                       (1, 2, 4096)    32.00K
Tensor497                                          (1, 2, 1)   512.00B
Tensor498                                       (1, 2, 4096)    32.00K
Tensor499                                       (1, 2, 4096)    32.00K
Tensor500                                      (1, 2, 11008)    86.00K
Tensor501                                      (1, 2, 11008)    86.00K
Tensor502                                      (1, 2, 11008)    86.00K
Tensor503                                      (1, 2, 11008)    86.00K
Tensor504                                       (1, 2, 4096)    32.00K
Tensor505                                          (1, 2, 1)   512.00B
Tensor506                                       (1, 2, 4096)    32.00K
Tensor507                                       (1, 2, 4096)    32.00K
Tensor508                                       (1, 2, 4096)    32.00K
Tensor509                                    (1, 32, 2, 128)     0.00B
Tensor510                                     (1, 1, 2, 128)     0.00B
Tensor511                                     (1, 1, 2, 128)     0.00B
Tensor512                                    (1, 32, 2, 128)    32.00K
Tensor513                                    (1, 32, 2, 128)    32.00K
Tensor514                                    (1, 32, 2, 128)    32.00K
Tensor515                                       (1, 2, 4096)    32.00K
Tensor516                                          (1, 2, 1)   512.00B
Tensor517                                       (1, 2, 4096)    32.00K
Tensor518                                       (1, 2, 4096)    32.00K
Tensor519                                      (1, 2, 11008)    86.00K
Tensor520                                      (1, 2, 11008)    86.00K
Tensor521                                      (1, 2, 11008)    86.00K
Tensor522                                      (1, 2, 11008)    86.00K
Tensor523                                       (1, 2, 4096)    32.00K
Tensor524                                          (1, 2, 1)   512.00B
Tensor525                                       (1, 2, 4096)    32.00K
Tensor526                                       (1, 2, 4096)    32.00K
Tensor527                                       (1, 2, 4096)    32.00K
Tensor528                                    (1, 32, 2, 128)     0.00B
Tensor529                                     (1, 1, 2, 128)     0.00B
Tensor530                                     (1, 1, 2, 128)     0.00B
Tensor531                                    (1, 32, 2, 128)    32.00K
Tensor532                                    (1, 32, 2, 128)    32.00K
Tensor533                                    (1, 32, 2, 128)    32.00K
Tensor534                                       (1, 2, 4096)    32.00K
Tensor535                                          (1, 2, 1)   512.00B
Tensor536                                       (1, 2, 4096)    32.00K
Tensor537                                       (1, 2, 4096)    32.00K
Tensor538                                      (1, 2, 11008)    86.00K
Tensor539                                      (1, 2, 11008)    86.00K
Tensor540                                      (1, 2, 11008)    86.00K
Tensor541                                      (1, 2, 11008)    86.00K
Tensor542                                       (1, 2, 4096)    32.00K
Tensor543                                          (1, 2, 1)   512.00B
Tensor544                                       (1, 2, 4096)    32.00K
Tensor545                                       (1, 2, 4096)    32.00K
Tensor546                                       (1, 2, 4096)    32.00K
Tensor547                                    (1, 32, 2, 128)     0.00B
Tensor548                                     (1, 1, 2, 128)     0.00B
Tensor549                                     (1, 1, 2, 128)     0.00B
```

```
Tensor550                                    (1, 32, 2, 128)    32.00K
Tensor551                                    (1, 32, 2, 128)    32.00K
Tensor552                                    (1, 32, 2, 128)    32.00K
Tensor553                                       (1, 2, 4096)    32.00K
Tensor554                                          (1, 2, 1)   512.00B
Tensor555                                       (1, 2, 4096)    32.00K
Tensor556                                       (1, 2, 4096)    32.00K
Tensor557                                      (1, 2, 11008)    86.00K
Tensor558                                      (1, 2, 11008)    86.00K
Tensor559                                      (1, 2, 11008)    86.00K
Tensor560                                      (1, 2, 11008)    86.00K
Tensor561                                       (1, 2, 4096)    32.00K
Tensor562                                          (1, 2, 1)   512.00B
Tensor563                                       (1, 2, 4096)    32.00K
Tensor564                                       (1, 2, 4096)    32.00K
Tensor565                                       (1, 2, 4096)    32.00K
Tensor566                                    (1, 32, 2, 128)     0.00B
Tensor567                                     (1, 1, 2, 128)     0.00B
Tensor568                                     (1, 1, 2, 128)     0.00B
Tensor569                                    (1, 32, 2, 128)    32.00K
Tensor570                                    (1, 32, 2, 128)    32.00K
Tensor571                                    (1, 32, 2, 128)    32.00K
Tensor572                                       (1, 2, 4096)    32.00K
Tensor573                                          (1, 2, 1)   512.00B
Tensor574                                       (1, 2, 4096)    32.00K
Tensor575                                       (1, 2, 4096)    32.00K
Tensor576                                      (1, 2, 11008)    86.00K
Tensor577                                      (1, 2, 11008)    86.00K
Tensor578                                      (1, 2, 11008)    86.00K
Tensor579                                      (1, 2, 11008)    86.00K
Tensor580                                       (1, 2, 4096)    32.00K
Tensor581                                          (1, 2, 1)   512.00B
Tensor582                                       (1, 2, 4096)    32.00K
Tensor583                                       (1, 2, 4096)    32.00K
Tensor584                                       (1, 2, 4096)    32.00K
Tensor585                                    (1, 32, 2, 128)     0.00B
Tensor586                                     (1, 1, 2, 128)     0.00B
Tensor587                                     (1, 1, 2, 128)     0.00B
Tensor588                                    (1, 32, 2, 128)    32.00K
Tensor589                                    (1, 32, 2, 128)    32.00K
Tensor590                                    (1, 32, 2, 128)    32.00K
Tensor591                                       (1, 2, 4096)    32.00K
Tensor592                                          (1, 2, 1)   512.00B
Tensor593                                       (1, 2, 4096)    32.00K
Tensor594                                       (1, 2, 4096)    32.00K
Tensor595                                      (1, 2, 11008)    86.00K
Tensor596                                      (1, 2, 11008)    86.00K
Tensor597                                      (1, 2, 11008)    86.00K
Tensor598                                      (1, 2, 11008)    86.00K
Tensor599                                       (1, 2, 4096)    32.00K
Tensor600                                          (1, 2, 1)   512.00B
Tensor601                                       (1, 2, 4096)    32.00K
Tensor602                                       (1, 2, 4096)    32.00K
Tensor603                                       (1, 2, 4096)    32.00K
Tensor604                                    (1, 32, 2, 128)     0.00B
Tensor605                                     (1, 1, 2, 128)     0.00B
Tensor606                                     (1, 1, 2, 128)     0.00B
Tensor607                                    (1, 32, 2, 128)    32.00K
Tensor608                                    (1, 32, 2, 128)    32.00K
Tensor609                                    (1, 32, 2, 128)    32.00K
Tensor610                                       (1, 2, 4096)    32.00K
Tensor611                                          (1, 2, 1)   512.00B
Tensor612                                       (1, 2, 4096)    32.00K
Tensor613                                       (1, 2, 4096)    32.00K
Tensor614                                        (1, 2, 128)     0.00B
Tensor615                                        (1, 2, 128)     0.00B
lm_head.weight                                 (32000, 4096)   500.00M
Tensor616                                              (64,)   512.00B
model.embed_tokens.weight                      (32000, 4096)   500.00M
model.norm.weight                                    (4096,)    16.00K
model.layers.0.input_layernorm.weight                (4096,)    16.00K
model.layers.0.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.1.input_layernorm.weight                (4096,)    16.00K
model.layers.1.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.2.input_layernorm.weight                (4096,)    16.00K
model.layers.2.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.3.input_layernorm.weight                (4096,)    16.00K
model.layers.3.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.4.input_layernorm.weight                (4096,)    16.00K
model.layers.4.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.5.input_layernorm.weight                (4096,)    16.00K
model.layers.5.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.6.input_layernorm.weight                (4096,)    16.00K
model.layers.6.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.7.input_layernorm.weight                (4096,)    16.00K
model.layers.7.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.8.input_layernorm.weight                (4096,)    16.00K
model.layers.8.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.9.input_layernorm.weight                (4096,)    16.00K
model.layers.9.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.10.input_layernorm.weight               (4096,)    16.00K
model.layers.10.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.11.input_layernorm.weight               (4096,)    16.00K
model.layers.11.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.12.input_layernorm.weight               (4096,)    16.00K
model.layers.12.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.13.input_layernorm.weight               (4096,)    16.00K
model.layers.13.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.14.input_layernorm.weight               (4096,)    16.00K
model.layers.14.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.15.input_layernorm.weight               (4096,)    16.00K
model.layers.15.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.16.input_layernorm.weight               (4096,)    16.00K
model.layers.16.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.17.input_layernorm.weight               (4096,)    16.00K
model.layers.17.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.18.input_layernorm.weight               (4096,)    16.00K
model.layers.18.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.19.input_layernorm.weight               (4096,)    16.00K
model.layers.19.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.20.input_layernorm.weight               (4096,)    16.00K
model.layers.20.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.21.input_layernorm.weight               (4096,)    16.00K
model.layers.21.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.22.input_layernorm.weight               (4096,)    16.00K
model.layers.22.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.23.input_layernorm.weight               (4096,)    16.00K
model.layers.23.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.24.input_layernorm.weight               (4096,)    16.00K
model.layers.24.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.25.input_layernorm.weight               (4096,)    16.00K
model.layers.25.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.26.input_layernorm.weight               (4096,)    16.00K
model.layers.26.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.27.input_layernorm.weight               (4096,)    16.00K
model.layers.27.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.28.input_layernorm.weight               (4096,)    16.00K
model.layers.28.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.29.input_layernorm.weight               (4096,)    16.00K
model.layers.29.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.30.input_layernorm.weight               (4096,)    16.00K
model.layers.30.post_attention_layernorm.weight             (4096,)    16.00K
model.layers.31.input_layernorm.weight               (4096,)    16.00K
model.layers.31.post_attention_layernorm.weight             (4096,)    16.00K
Tensor617                                              (64,)   512.00B
Tensor618                                              (64,)   512.00B
Tensor619                                              (64,)   512.00B
Tensor620                                              (64,)   512.00B
Tensor621                                              (64,)   512.00B
Tensor622                                              (64,)   512.00B
Tensor623                                              (64,)   512.00B
Tensor624                                              (64,)   512.00B
Tensor625                                              (64,)   512.00B
Tensor626                                              (64,)   512.00B
Tensor627                                              (64,)   512.00B
Tensor628                                              (64,)   512.00B
Tensor629                                              (64,)   512.00B
Tensor630                                              (64,)   512.00B
Tensor631                                              (64,)   512.00B
Tensor632                                              (64,)   512.00B
Tensor633                                              (64,)   512.00B
Tensor634                                              (64,)   512.00B
Tensor635                                              (64,)   512.00B
Tensor636                                              (64,)   512.00B
Tensor637                                              (64,)   512.00B
Tensor638                                              (64,)   512.00B
Tensor639                                              (64,)   512.00B
Tensor640                                              (64,)   512.00B
Tensor641                                              (64,)   512.00B
Tensor642                                              (64,)   512.00B
Tensor643                                              (64,)   512.00B
Tensor644                                              (64,)   512.00B
Tensor645                                              (64,)   512.00B
Tensor646                                              (64,)   512.00B
Tensor647                                              (64,)   512.00B
Tensor648                                              (64,)   512.00B
```

```
model.layers.0.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.0.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.0.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.0.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.0.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.0.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.0.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.1.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.1.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.1.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.1.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.1.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.1.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.1.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.2.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.2.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.2.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.2.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.2.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.2.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.2.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.3.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.3.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.3.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.3.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.3.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.3.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.3.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.4.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.4.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.4.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.4.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.4.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.4.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.4.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.5.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.5.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.5.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.5.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.5.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.5.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.5.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.6.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.6.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.6.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.6.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.6.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.6.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.6.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.7.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.7.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.7.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.7.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.7.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.7.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.7.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.8.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.8.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.8.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.8.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.8.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.8.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.8.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.9.self_attn.q_proj.weight          (4096, 4096)    64.00M
model.layers.9.self_attn.k_proj.weight          (4096, 4096)    64.00M
model.layers.9.self_attn.v_proj.weight          (4096, 4096)    64.00M
model.layers.9.self_attn.o_proj.weight          (4096, 4096)    64.00M
model.layers.9.mlp.gate_proj.weight            (11008, 4096)   172.00M
model.layers.9.mlp.up_proj.weight              (11008, 4096)   172.00M
model.layers.9.mlp.down_proj.weight            (4096, 11008)   172.00M
model.layers.10.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.10.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.10.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.10.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.10.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.10.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.10.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.11.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.11.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.11.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.11.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.11.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.11.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.11.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.12.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.12.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.12.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.12.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.12.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.12.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.12.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.13.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.13.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.13.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.13.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.13.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.13.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.13.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.14.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.14.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.14.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.14.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.14.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.14.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.14.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.15.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.15.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.15.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.15.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.15.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.15.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.15.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.16.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.16.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.16.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.16.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.16.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.16.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.16.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.17.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.17.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.17.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.17.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.17.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.17.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.17.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.18.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.18.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.18.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.18.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.18.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.18.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.18.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.19.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.19.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.19.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.19.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.19.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.19.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.19.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.20.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.20.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.20.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.20.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.20.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.20.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.20.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.21.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.21.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.21.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.21.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.21.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.21.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.21.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.22.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.22.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.22.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.22.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.22.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.22.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.22.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.23.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.23.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.23.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.23.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.23.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.23.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.23.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.24.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.24.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.24.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.24.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.24.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.24.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.24.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.25.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.25.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.25.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.25.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.25.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.25.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.25.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.26.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.26.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.26.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.26.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.26.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.26.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.26.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.27.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.27.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.27.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.27.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.27.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.27.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.27.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.28.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.28.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.28.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.28.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.28.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.28.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.28.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.29.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.29.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.29.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.29.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.29.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.29.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.29.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.30.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.30.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.30.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.30.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.30.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.30.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.30.mlp.down_proj.weight           (4096, 11008)   172.00M
model.layers.31.self_attn.q_proj.weight         (4096, 4096)    64.00M
model.layers.31.self_attn.k_proj.weight         (4096, 4096)    64.00M
model.layers.31.self_attn.v_proj.weight         (4096, 4096)    64.00M
model.layers.31.self_attn.o_proj.weight         (4096, 4096)    64.00M
model.layers.31.mlp.gate_proj.weight           (11008, 4096)   172.00M
model.layers.31.mlp.up_proj.weight             (11008, 4096)   172.00M
model.layers.31.mlp.down_proj.weight           (4096, 11008)   172.00M
-------------------------------------------------------------------------------
Total Tensors: 6744224964       Used Memory: 25.12G
-------------------------------------------------------------------------------
```

```
C:\Users\xxc13\anaconda3\Lib\site-packages\pytorch_memlab\mem_reporter.py:65: FutureWarning: `torch.distributed.
reduce_op` is deprecated, please use `torch.distributed.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: UserWarning: TypedStorage is depre
cated. It will be removed in the future and UntypedStorage will be the only storage class. This should only matt
er to you if you are using storages directly.  To access UntypedStorage directly, use tensor.untyped_storage() i
nstead of tensor.storage()
  fact_numel = tensor.storage().size()
```

In [3]:

```
# 清除前向传播的变量，确保不占用内存
del outputs, loss
gc.collect()  # 强制清理内存
```

Out[3]:

```
46
```

In [ ]: