我的结构：

|  |
| --- |
| Transformer(  (encoder): Encoder(  (word\_embedding): Embedding(10046, 512)  (position\_embedding): Embedding(25, 512)  (layer): ModuleList(  (0-5): 6 x TransformerBlock(  (attention): MultiHeadAttention(  (values): Linear(in\_features=64, out\_features=64, bias=False)  (keys): Linear(in\_features=64, out\_features=64, bias=False)  (queries): Linear(in\_features=64, out\_features=64, bias=False)  (fc\_out): Linear(in\_features=512, out\_features=512, bias=True)  )  (norm1): LayerNorm((512,), eps=1e-05, elementwise\_affine=True)  (norm2): LayerNorm((512,), eps=1e-05, elementwise\_affine=True)  (feed\_forward): Sequential(  (0): Linear(in\_features=512, out\_features=2048, bias=True)  (1): ReLU()  (2): Linear(in\_features=2048, out\_features=512, bias=True)  )  (dropout): Dropout(p=0.1, inplace=False)  )  )  (dropout): Dropout(p=0.1, inplace=False)  )  (decoder): Decoder(  (word\_embedding): Embedding(6749, 512)  (position\_embedding): Embedding(25, 512)  (layers): ModuleList(  (0-5): 6 x DecoderBlock(  (attention): MultiHeadAttention(  (values): Linear(in\_features=64, out\_features=64, bias=False)  (keys): Linear(in\_features=64, out\_features=64, bias=False)  (queries): Linear(in\_features=64, out\_features=64, bias=False)  (fc\_out): Linear(in\_features=512, out\_features=512, bias=True)  )  (norm): LayerNorm((512,), eps=1e-05, elementwise\_affine=True)  (transformer\_block): TransformerBlock(  (attention): MultiHeadAttention(  (values): Linear(in\_features=64, out\_features=64, bias=False)  (keys): Linear(in\_features=64, out\_features=64, bias=False)  (queries): Linear(in\_features=64, out\_features=64, bias=False)  (fc\_out): Linear(in\_features=512, out\_features=512, bias=True)  )  (norm1): LayerNorm((512,), eps=1e-05, elementwise\_affine=True)  (norm2): LayerNorm((512,), eps=1e-05, elementwise\_affine=True)  (feed\_forward): Sequential(  (0): Linear(in\_features=512, out\_features=2048, bias=True)  (1): ReLU()  (2): Linear(in\_features=2048, out\_features=512, bias=True)  )  (dropout): Dropout(p=0.1, inplace=False)  )  (dropout): Dropout(p=0.1, inplace=False)  )  )  (fc\_out): Linear(in\_features=512, out\_features=6749, bias=True)  (dropout): Dropout(p=0.1, inplace=False)  )  ) |

1. **SnapshotAPI**

**1.介绍：**

在PyTorch2.1中的显存snapshot功能被增强（在1.X里面也有snapshot的记录操作），可以将显存消耗可视化，特点是查看简单、更易理解。

调用步骤如下：

|  |
| --- |
| # 开始（训练/推理前）:  torch.cuda.memory.\_record\_memory\_history(max\_entries=80000)  #保存（迭代结束后）:  torch.cuda.memory.\_dump\_snapshot(file\_name)  #停止（分析完成）:  torch.cuda.memory.\_record\_memory\_history(enabled=None) |

注：max\_entries：最多使用多少个alloc/free events来记录内存开销，内存的操作需要用events来记录。当记录数据溢出时，系统只保留最后的max\_entries个events量的数据。参数设置过小保存数据会不足，设置过大可能会影响运行

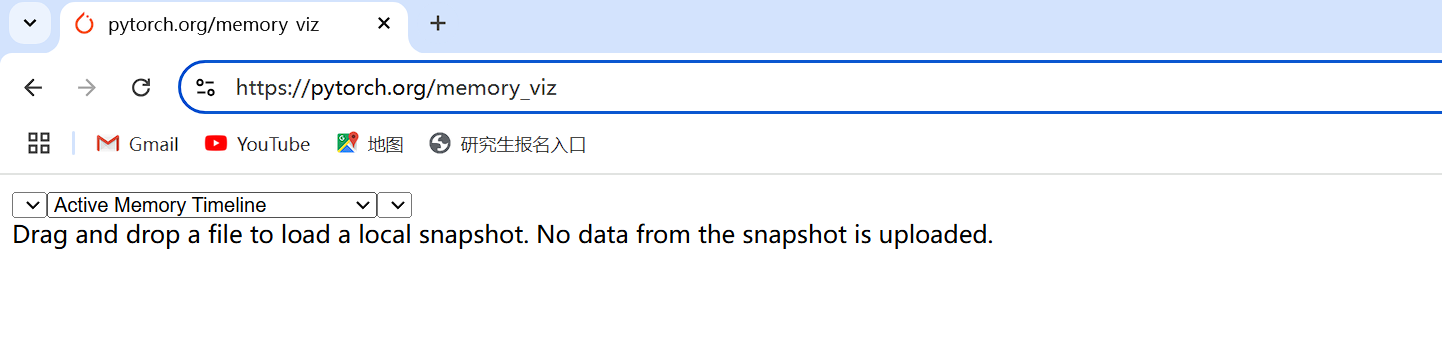
context： 选择需要跟踪的数据类型[None,"state","alloc","all"]。"state"是指记录当前使用内存情况，“alloc"是指通过alloc调用过的内存跟踪，（如果不会设置按默认直即可），all缺省。

**2. 实际调用：**

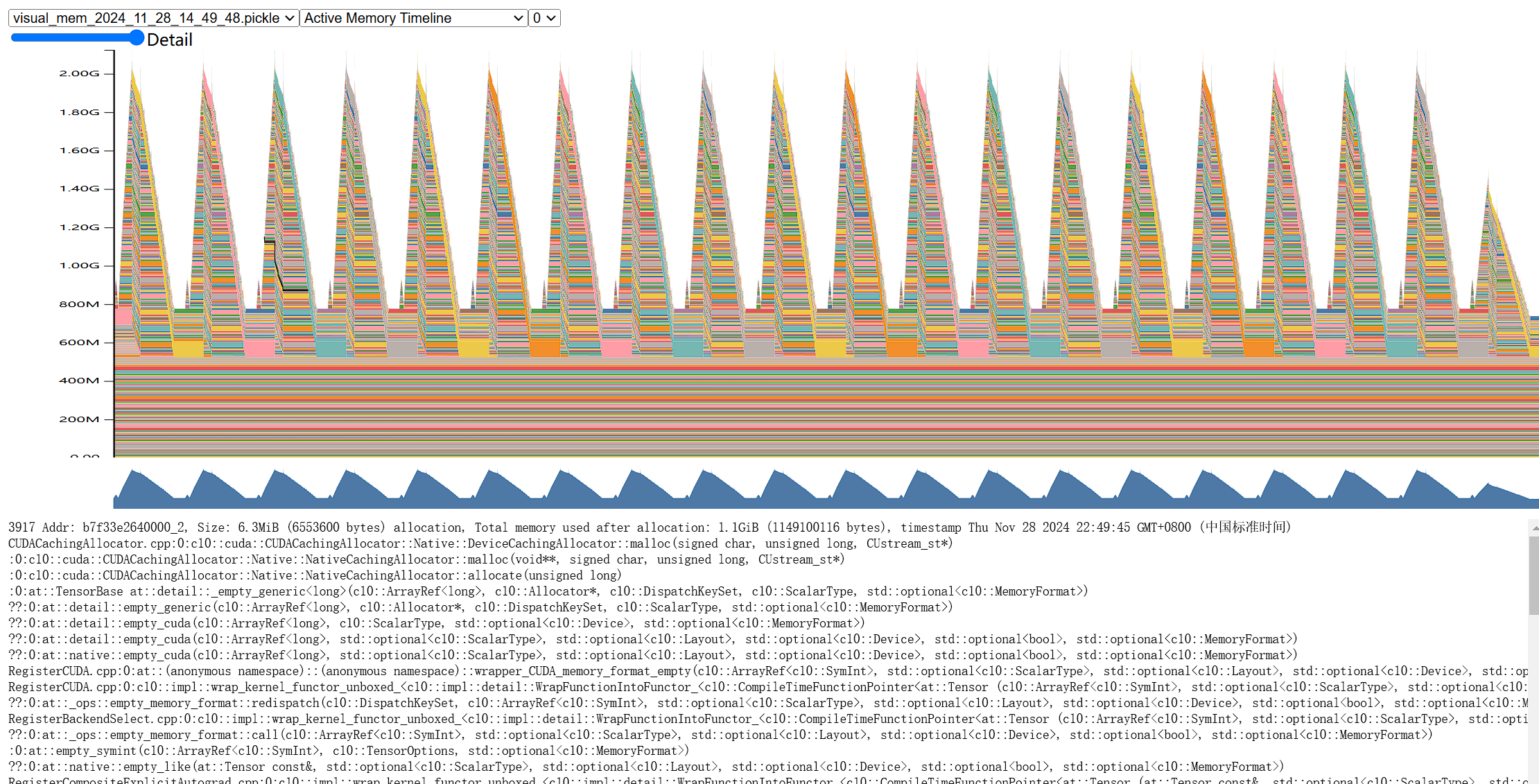
|  |
| --- |
| # Start recording memory snapshot history      torch.cuda.memory.\_record\_memory\_history(*max\_entries*=100000)  train\_model(model, data\_loader, num\_epochs, device, optimizer, criterion, save\_path)      timestamp = datetime.now().strftime('%Y\_%m\_%d\_%H\_%M\_%S')      file\_name = *f*"visual\_mem\_{timestamp}.pickle"      # save record      torch.cuda.memory.\_dump\_snapshot(file\_name)      # Stop recording memory snapshot history      torch.cuda.memory.\_record\_memory\_history(*enabled*=None) |

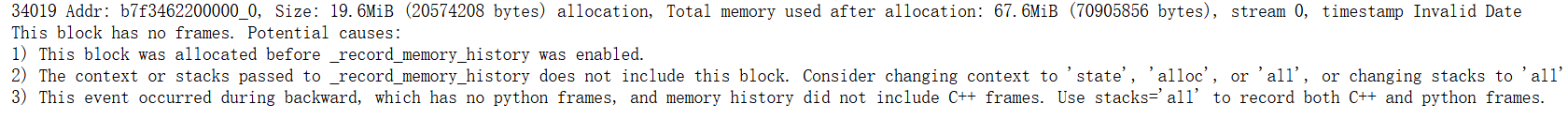
**3. 可视化运行结果**：

首先将上述代码生成的pickle文件拖拽到浏览器（如chrome等）的下面网址中:*https://pytorch.org/memory\_viz* （需要挂梯子）



Batch\_size=128可视化结果如下：

注意到底下长方形一部分内存分配描述是这样的：



表示在记录snapshot历史前就已经分配了。

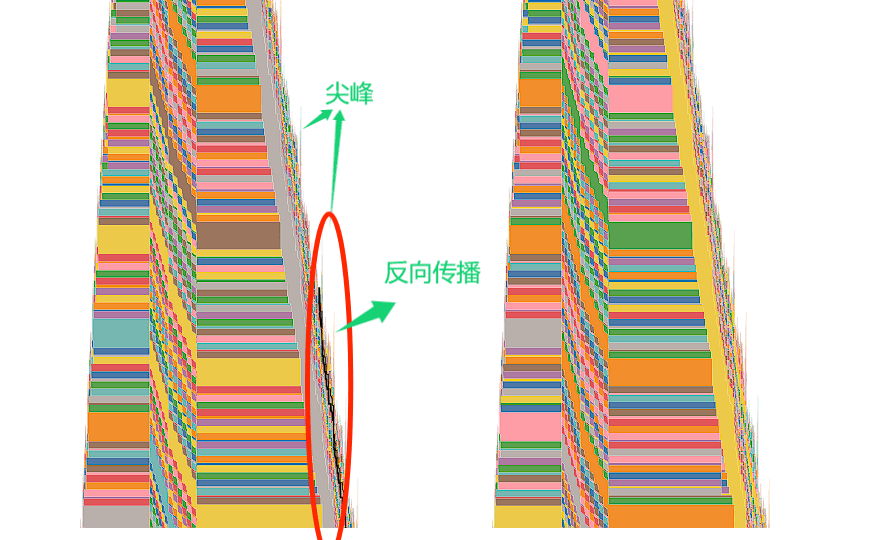
**4. snapshot数据分析：**

（1）激活内存数据(Active Memory Timeline)：

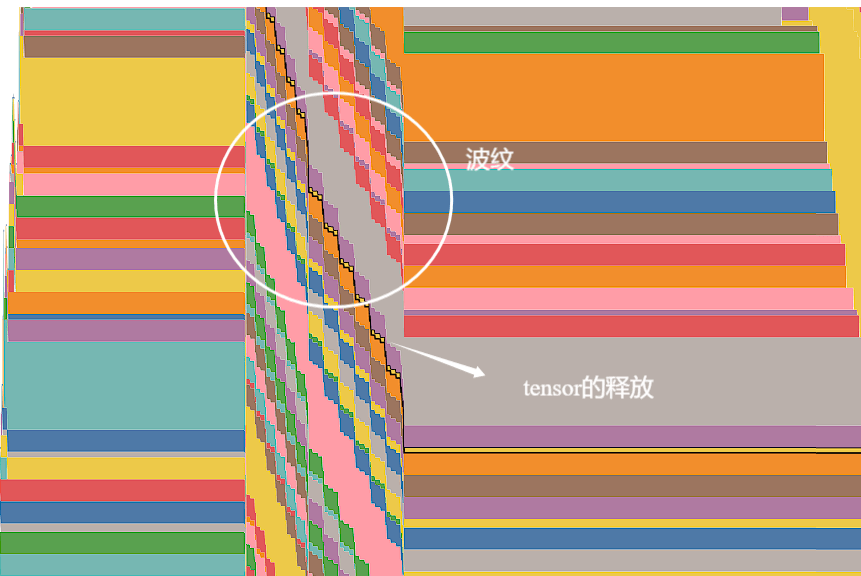
这部分数据主要是记录张量tensor在计算过程中占用的内存以及其存活的周期(在计算中从创建到销毁的时间)；同时，能够查看每个tensor的内存消耗调用堆栈（Python/C++），帮助理解哪部分代码调用导致了内存消耗。

数据分析：

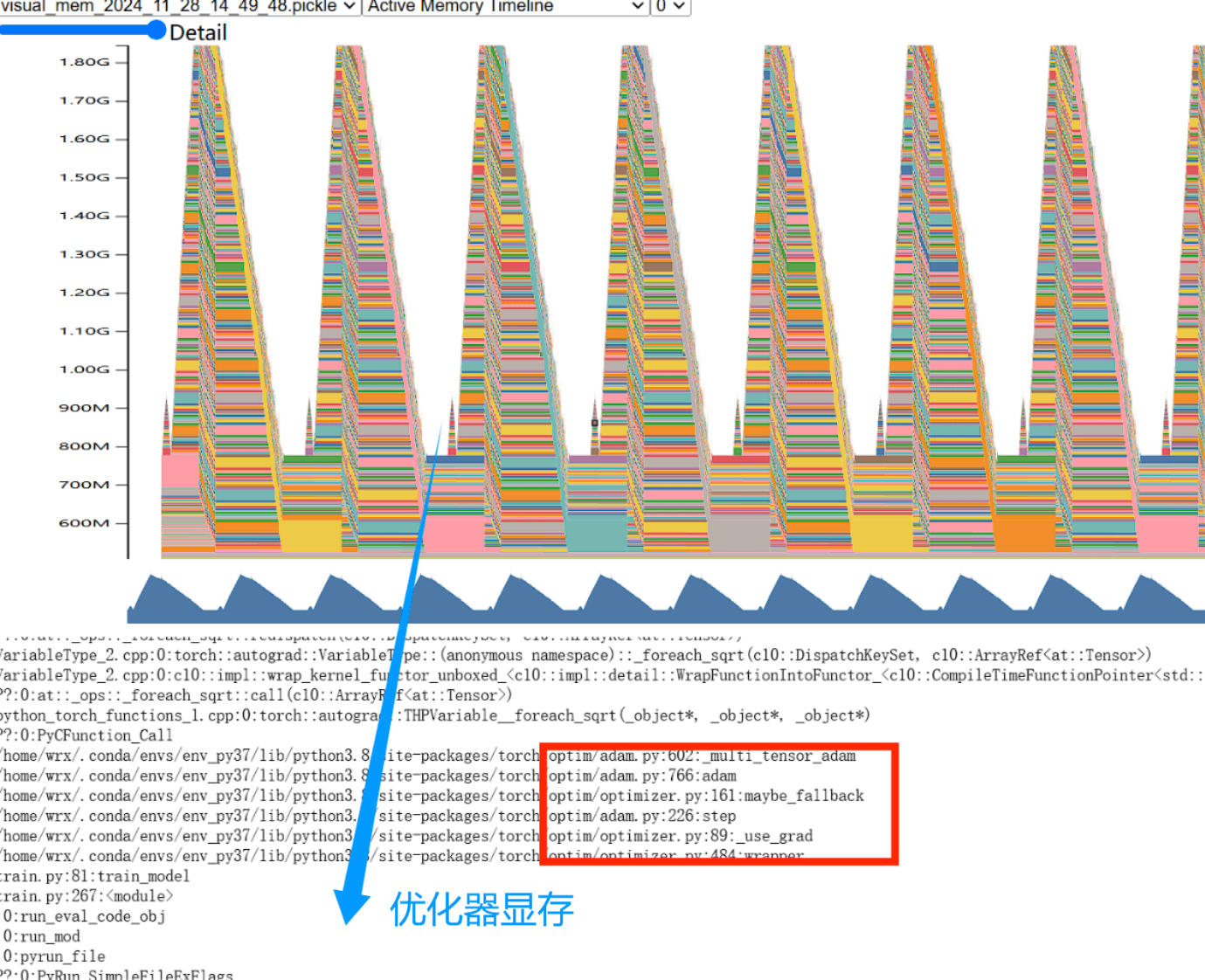
1. 反向传播部分/尖峰：后向计算时autograd(PyTorch 中的一个自动微分库,用来计算梯度)会产生额外的内存消耗，能够形成一些尖峰值。



1. 波纹：出现条形上行/下形波纹是因为其它的tensor被创建/释放了导致的显示绘图的变化，并非当前tensor占用的显存大小发生了改变，例如某些中间计算结果的 tensor 可能只在一次前向传播或反向传播中存在之后就被销毁，单个tensor数据消耗显存在生命周期内是保持不变的。

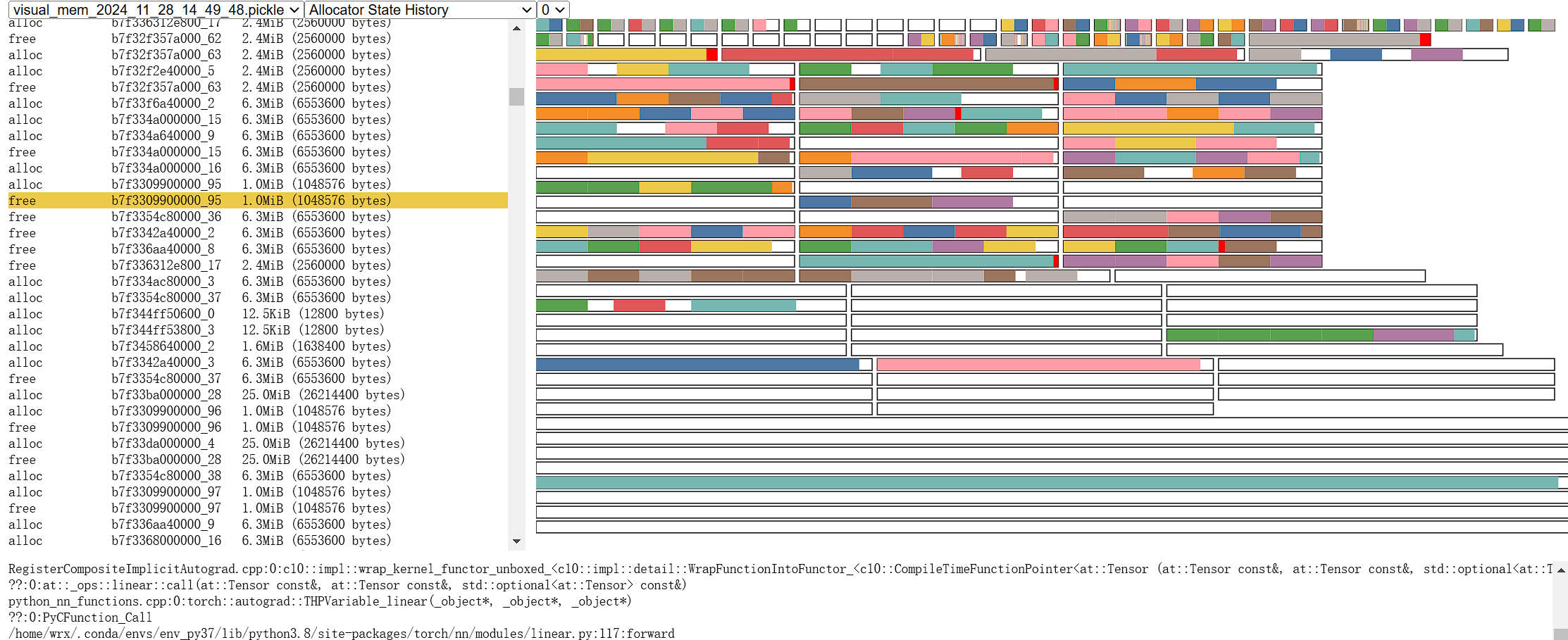


1. 优化器显存：



（2）内存块的使用与释放(Allocator State Hisotry)：

该图主要展示从CUDA里面申请的内存块是如何被alloc分配成小的block，以及这些小的block什么时候被释放掉了。(点击左侧地址块即可显示)如果把光标放置到这些彩色的block上面，能够跟踪到是什么操作在使用该显存地址。



例如：表示分配了地址b7f32f357a000\_65，大小为2.4MiB的块

|  |
| --- |
| (1.0GiB (1076242516 bytes) allocated / 2.3GiB (2512388096 bytes) reserved)  alloc b7f32f357a000\_65 2.4MiB (2560000 bytes) |

**二、 Profiler中显存可视化使用**

**1.介绍**：

PyTorch Profiler已支持把显存的snapshot数据记录到profiling中，通过prof.export\_memory\_timeline函数可导出snapshot数据。由于Profiler能够给一些数据打标签，所以可更加详细的记录具体是由哪个过程消耗了显存。通过可视化图表能直观的看到激活值、优化器、输入等操作的显存消耗，相比snapshot图表更加容易读懂，但没有snapshot数据那么精细。

**2. API调用步骤：**

开启方式：

|  |
| --- |
| with torch.profiler.profile(profile\_memory=True,with\_stack=True,on\_trace\_ready=trace\_handler,)： |

注：profile\_memory - 是否记录 tensor memory allocation/deallocation.

with\_stack (bool) – 追踪操作的调用堆栈.

on\_trace\_ready(Callable) – 记录完成后的后处理回调函数。

**3.实际调用**：

（1）导出内存消耗图表：

|  |
| --- |
| *def* trace\_handler(*prof*: torch.profiler.profile):     # 获取时间用于文件命名     timestamp = datetime.now().strftime('%Y\_%m\_%d\_%H\_%M\_%S')     file\_name = *f*"visual\_mem\_{timestamp}"     # 导出tracing格式的profiling  *prof*.export\_chrome\_trace(*f*"{file\_name}.json")     # 导出mem消耗可视化数据  *prof*.export\_memory\_timeline(*f*"{file\_name}.html", *device*="cuda:0") |

（2）修改train\_model函数

|  |
| --- |
| # 启动 PyTorch Profiler 来收集性能数据      with torch.profiler.profile(  *activities*=[torch.profiler.ProfilerActivity.CPU, torch.profiler.ProfilerActivity.CUDA],  *schedule*=torch.profiler.schedule(*wait*=0, *warmup*=0, *active*=6, *repeat*=2),  *record\_shapes*=True,  *profile\_memory*=True,  *with\_stack*=True,  *on\_trace\_ready*=trace\_handler,  # 训练完成后处理并导出 trace 数据      ) as prof:  …… |

Active = 6，生成6步的。Repeat=2，重复两次。

训练一轮就有多步了，就看这个步指的是一批次，一批次的训练需要经过完整的transformer模型。因此一步是一batch\_size的完整6层显存占用，因为6层(6encoder+6decoder)过程中，显存会不断增大的，(当然也会有一些不用的被释放，详细可用snapshot查看)完整的网络走完后，才会更新梯度和权重，进而继续训练。

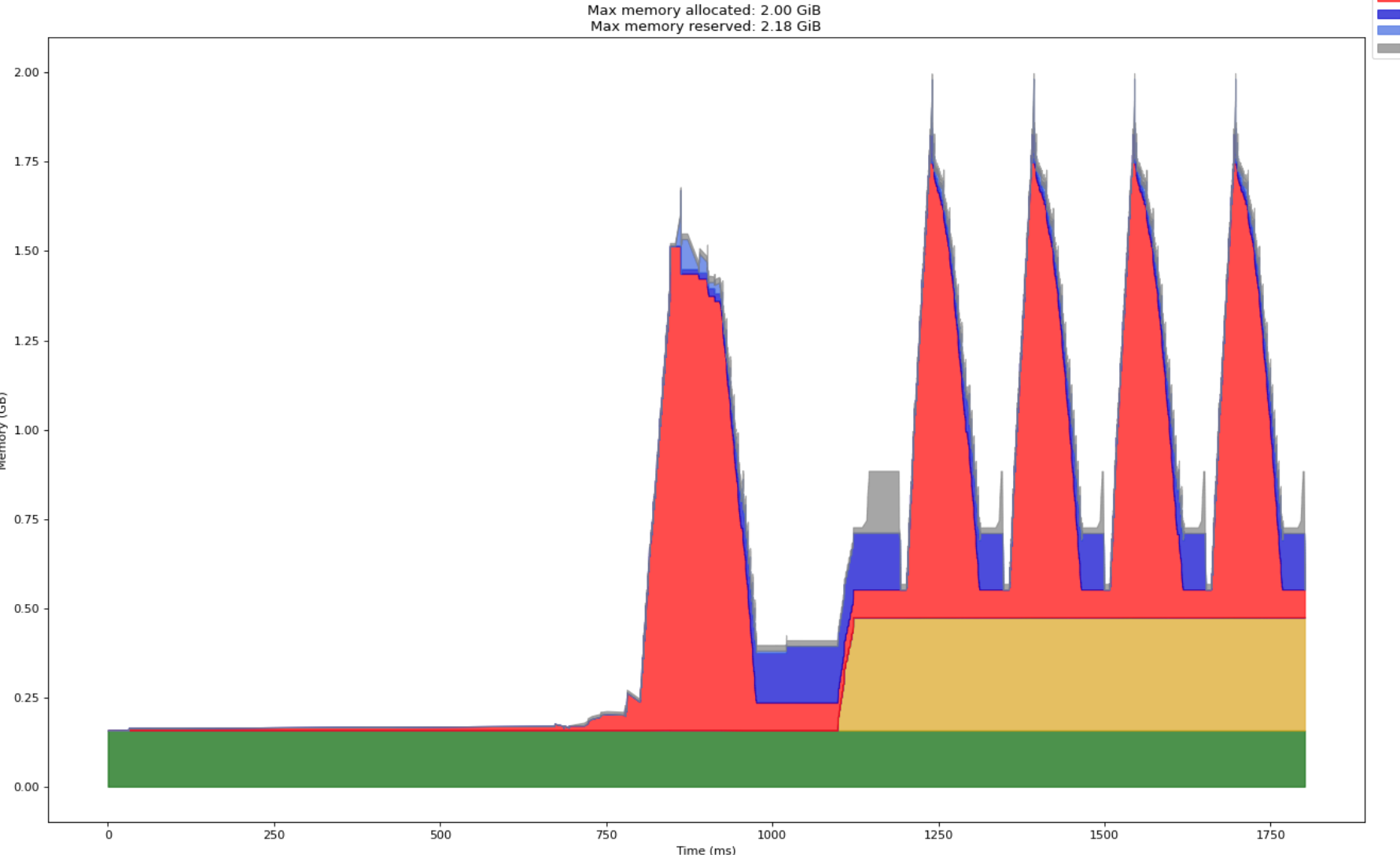
**4. 可视化结果（batch\_size=128）**：

横坐标为时间，纵坐标为显存占用，以下先说明各个颜色含义。

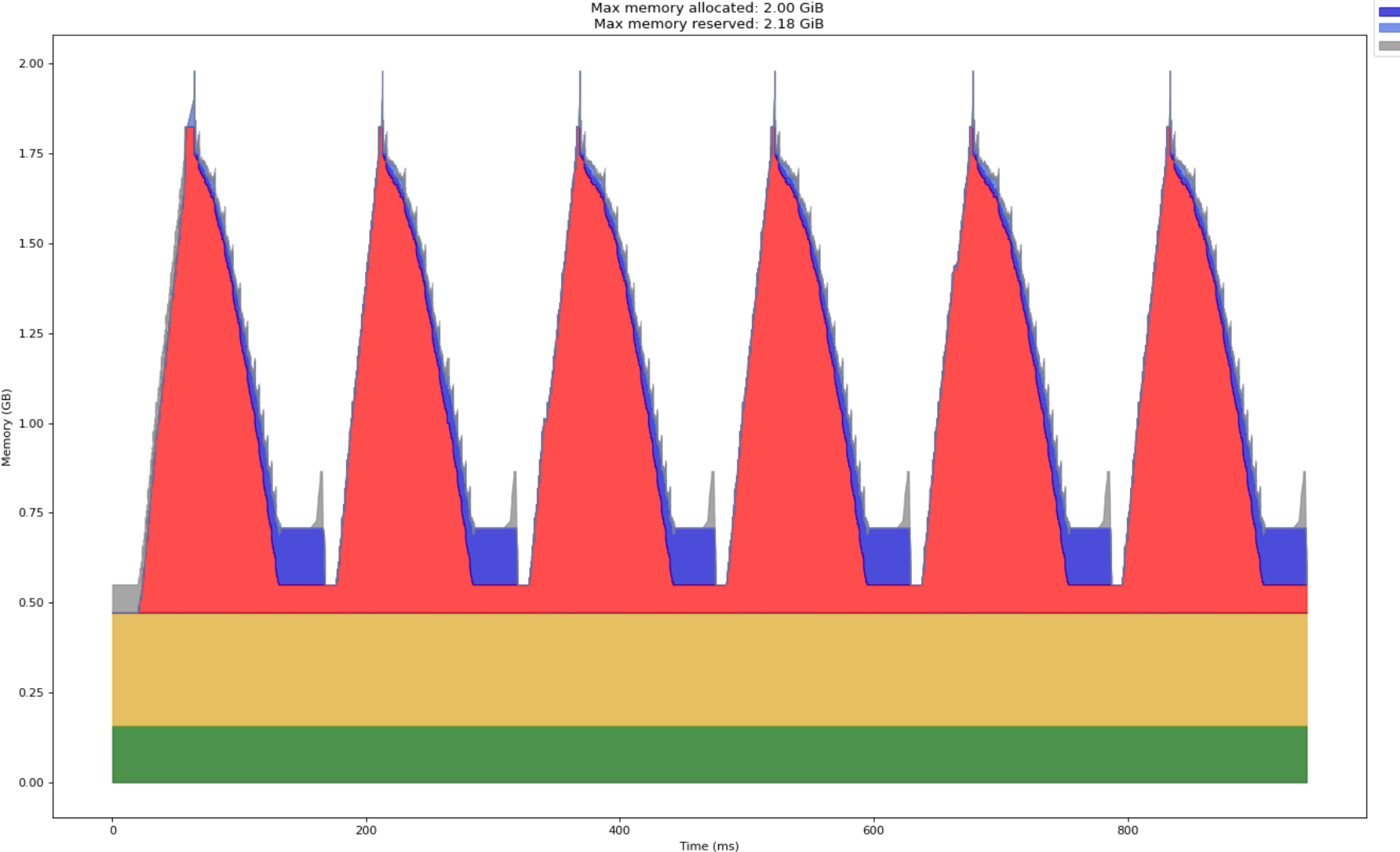
|  |
| --- |
| * parameter：模型参数 * optimizer\_state: 优化器状态参数 * input：输入值 * temporary：临时变量； * activation：激活值（前向运算产生） * gradient：梯度值 * autograd\_detail: 自动梯度产生的显存 * unknown：无法分类的消耗 |



前6步：



后六步：



可以看到最大显存占用表示的是尖峰值(后向计算时autograd会产生额外的内存消耗，能够形成一些尖峰值)。

Batch\_size=32:

前6步：



后6步：



综上，优化器状态和参数的显存占用基本不变，主要是激活值和梯度的显存占用在变化。

**三、transformer每层的显存使用**

1.方法：

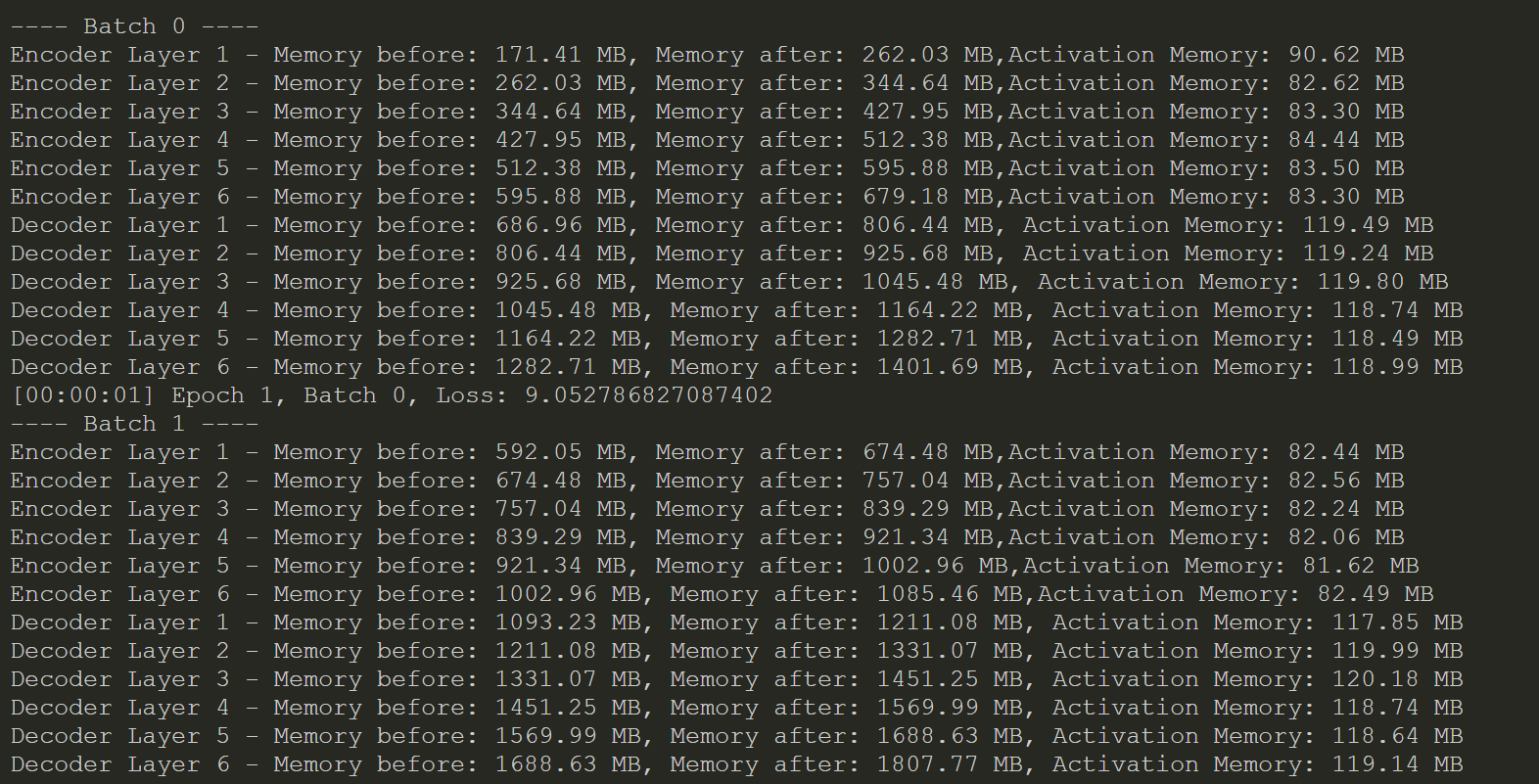
修改model.py，由于6层transformer = 6\*encoder + 6 \* decoder，所以就看这12层的激活值显存即可(参数和优化器状态显存大小基本不变)

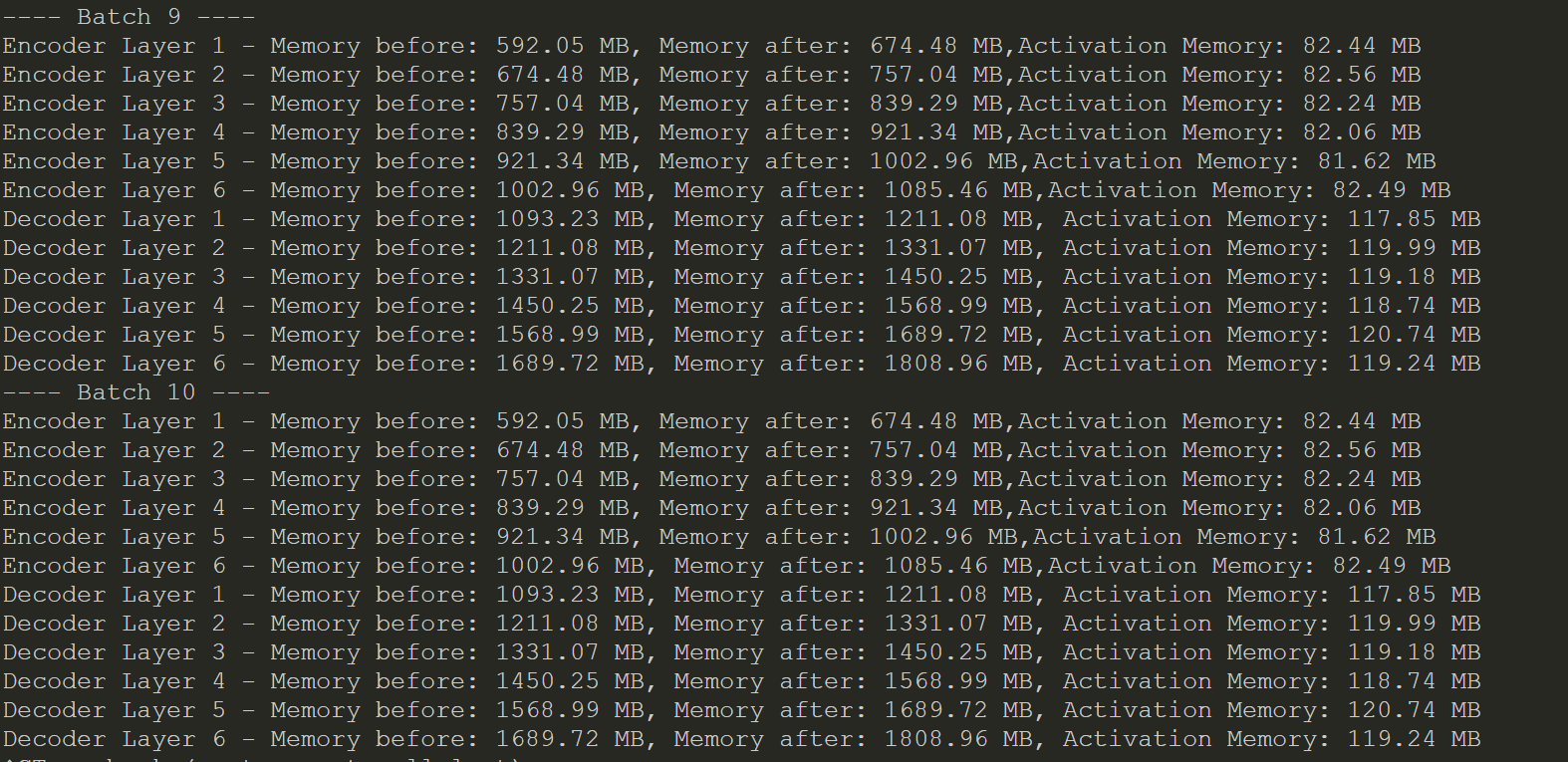
给encoder6层循环中的前向传播部分(不是反向传播)添加如下代码，捕获前后显存占用差值，decoder类似：

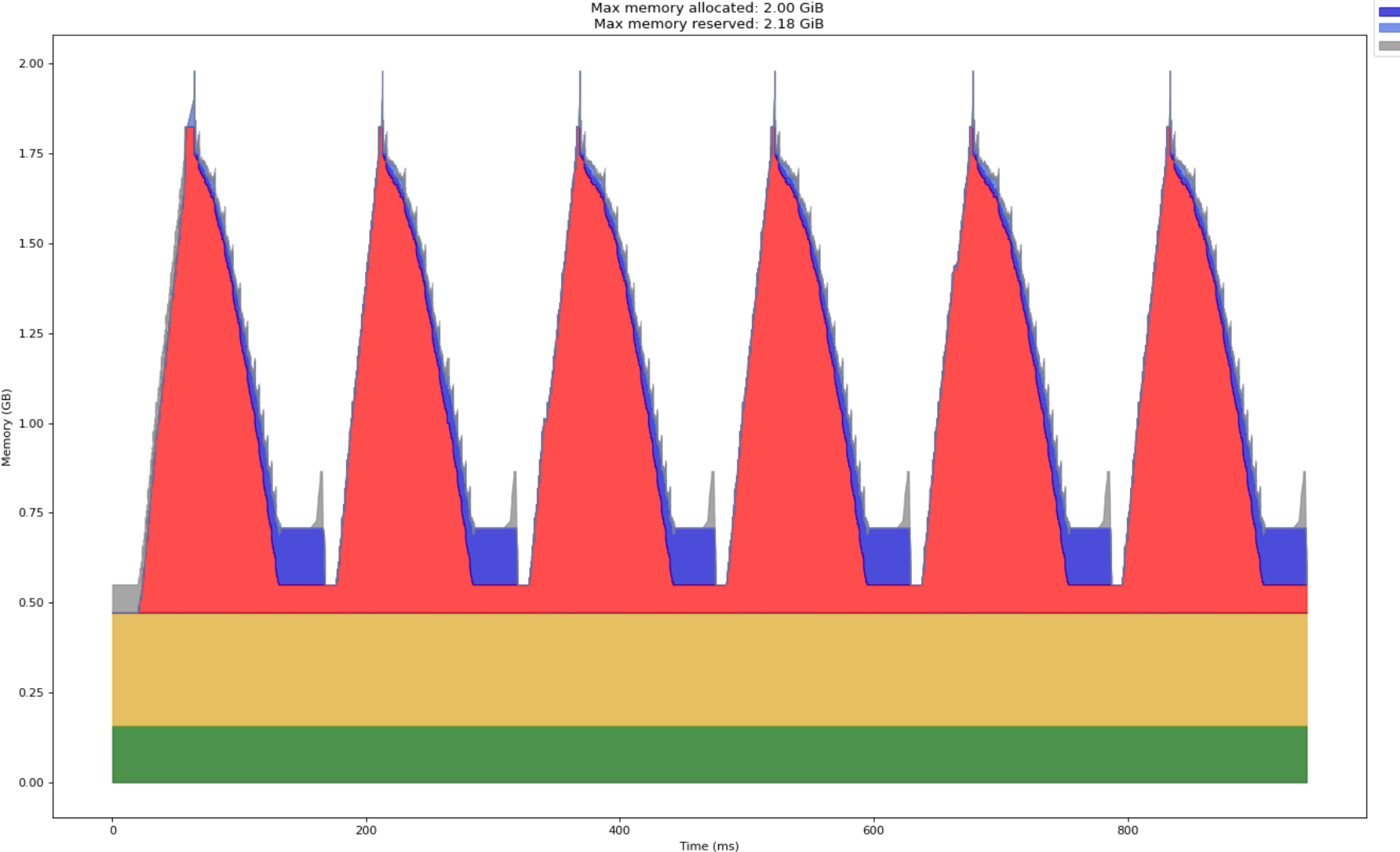
|  |
| --- |
| for i, layer in enumerate(*self*.layer):              # 每经过一层记录显存变化              allocated\_memory\_before = torch.cuda.memory\_allocated(*self*.device)              out = layer(out, out, out, *mask*)              allocated\_memory\_after = torch.cuda.memory\_allocated(*self*.device)              activation\_memory = (allocated\_memory\_after - allocated\_memory\_before) / 1024 \*\* 2  # 转换为 MB              print(*f*"Encoder Layer {i+1} - Memory before: {allocated\_memory\_before / 1024 \*\* 2*:.2f*} MB, "  *f*"Memory after: {allocated\_memory\_after / 1024 \*\* 2*:.2f*} MB,"  *f*"Activation Memory: {activation\_memory*:.2f*} MB")          return out |

2.运行结果：

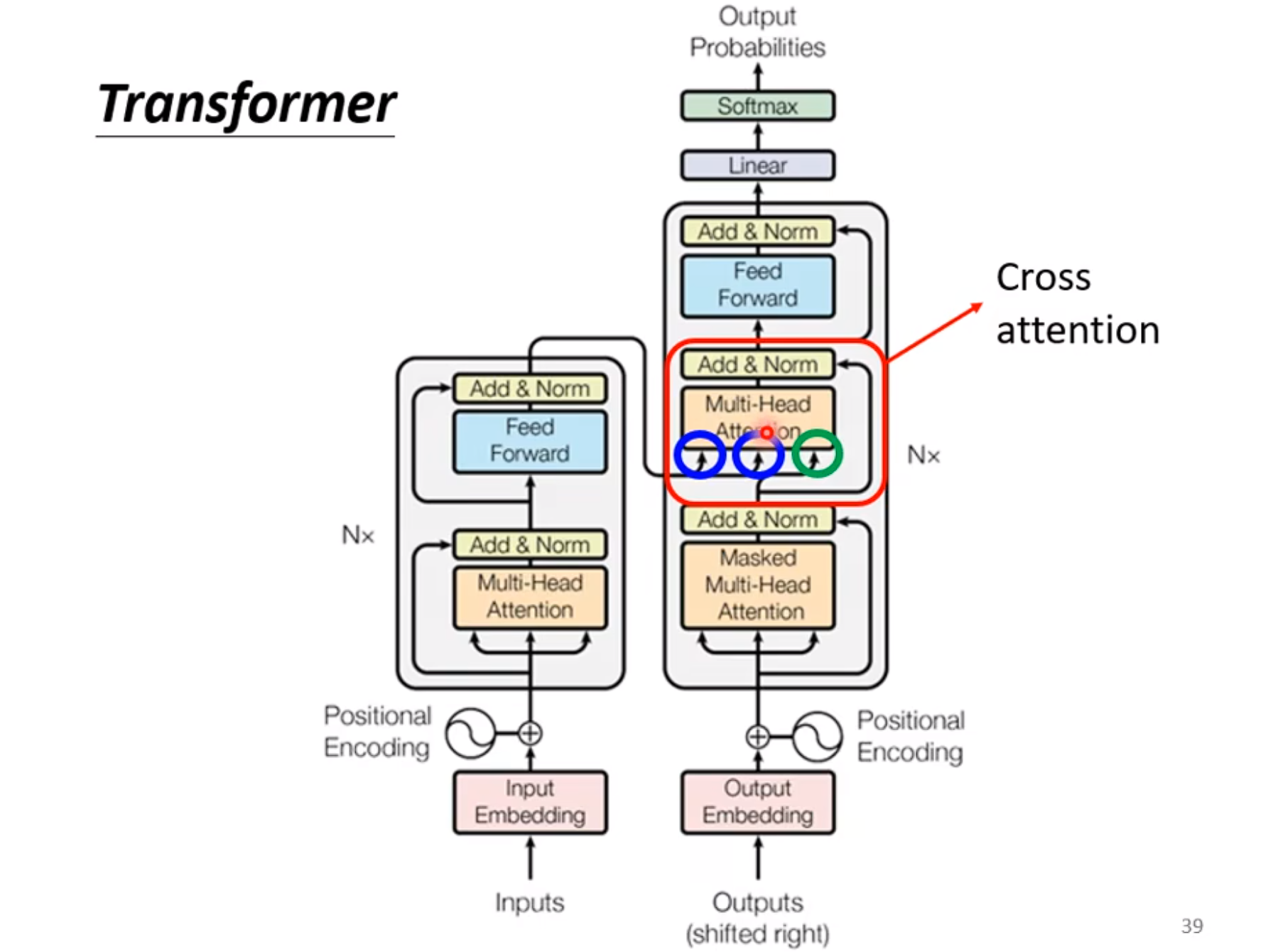
训练1轮次，查看每批次中每层结果：

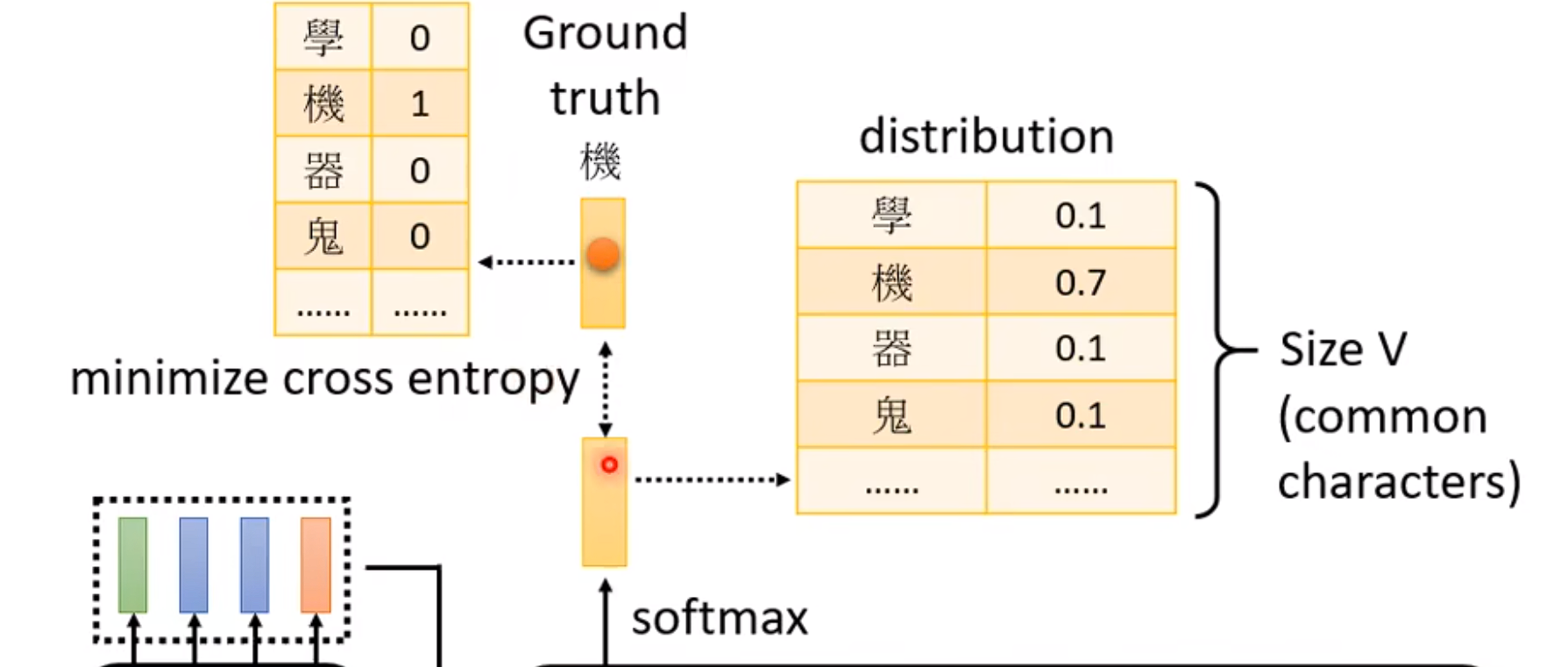






分析：相比于encoder，decoder不仅要jinx1自身对目标序列的注意力计算还多一个cross-attention层(encoder-decoder的资讯传递)，来计算目标序列与源序列之间的交叉注意力，这需要额外的计算和显存。此外，在训练过程中，Decoder 需要为每个目标序列的词预测一个概率分布（通常是词汇表的大小），因此 Decoder 需要存储更多的信息来计算每个时刻的词预测。





**四、transformer每层的参数使用与数目**：

1.方法：

|  |
| --- |
| *def* print\_layer\_params(*model*):      total\_params = 0      for name, param in *model*.named\_parameters():          num\_params = param.numel()          total\_params += num\_params          print(*f*"{name}: {num\_params}个参数")      print(*f*"总参数数量: {total\_params}个\n") |

2.运行结果：

|  |
| --- |
| 模型每层的参数数量：  encoder.word\_embedding.weight: 5143552个参数  encoder.position\_embedding.weight: 12800个参数  encoder.layer.0.attention.values.weight: 4096个参数  encoder.layer.0.attention.keys.weight: 4096个参数  encoder.layer.0.attention.queries.weight: 4096个参数  encoder.layer.0.attention.fc\_out.weight: 262144个参数  encoder.layer.0.attention.fc\_out.bias: 512个参数  encoder.layer.0.norm1.weight: 512个参数  encoder.layer.0.norm1.bias: 512个参数  encoder.layer.0.norm2.weight: 512个参数  encoder.layer.0.norm2.bias: 512个参数  encoder.layer.0.feed\_forward.0.weight: 1048576个参数  encoder.layer.0.feed\_forward.0.bias: 2048个参数  encoder.layer.0.feed\_forward.2.weight: 1048576个参数  encoder.layer.0.feed\_forward.2.bias: 512个参数  encoder.layer.1.attention.values.weight: 4096个参数  encoder.layer.1.attention.keys.weight: 4096个参数  encoder.layer.1.attention.queries.weight: 4096个参数  encoder.layer.1.attention.fc\_out.weight: 262144个参数  encoder.layer.1.attention.fc\_out.bias: 512个参数  encoder.layer.1.norm1.weight: 512个参数  encoder.layer.1.norm1.bias: 512个参数  encoder.layer.1.norm2.weight: 512个参数  encoder.layer.1.norm2.bias: 512个参数  encoder.layer.1.feed\_forward.0.weight: 1048576个参数  encoder.layer.1.feed\_forward.0.bias: 2048个参数  encoder.layer.1.feed\_forward.2.weight: 1048576个参数  encoder.layer.1.feed\_forward.2.bias: 512个参数  encoder.layer.2.attention.values.weight: 4096个参数  encoder.layer.2.attention.keys.weight: 4096个参数  encoder.layer.2.attention.queries.weight: 4096个参数  encoder.layer.2.attention.fc\_out.weight: 262144个参数  encoder.layer.2.attention.fc\_out.bias: 512个参数  encoder.layer.2.norm1.weight: 512个参数  encoder.layer.2.norm1.bias: 512个参数  encoder.layer.2.norm2.weight: 512个参数  encoder.layer.2.norm2.bias: 512个参数  encoder.layer.2.feed\_forward.0.weight: 1048576个参数  encoder.layer.2.feed\_forward.0.bias: 2048个参数  encoder.layer.2.feed\_forward.2.weight: 1048576个参数  encoder.layer.2.feed\_forward.2.bias: 512个参数  encoder.layer.3.attention.values.weight: 4096个参数  encoder.layer.3.attention.keys.weight: 4096个参数  encoder.layer.3.attention.queries.weight: 4096个参数  encoder.layer.3.attention.fc\_out.weight: 262144个参数  encoder.layer.3.attention.fc\_out.bias: 512个参数  encoder.layer.3.norm1.weight: 512个参数  encoder.layer.3.norm1.bias: 512个参数  encoder.layer.3.norm2.weight: 512个参数  encoder.layer.3.norm2.bias: 512个参数  encoder.layer.3.feed\_forward.0.weight: 1048576个参数  encoder.layer.3.feed\_forward.0.bias: 2048个参数  encoder.layer.3.feed\_forward.2.weight: 1048576个参数  encoder.layer.3.feed\_forward.2.bias: 512个参数  encoder.layer.4.attention.values.weight: 4096个参数  encoder.layer.4.attention.keys.weight: 4096个参数  encoder.layer.4.attention.queries.weight: 4096个参数  encoder.layer.4.attention.fc\_out.weight: 262144个参数  encoder.layer.4.attention.fc\_out.bias: 512个参数  encoder.layer.4.norm1.weight: 512个参数  encoder.layer.4.norm1.bias: 512个参数  encoder.layer.4.norm2.weight: 512个参数  encoder.layer.4.norm2.bias: 512个参数  encoder.layer.4.feed\_forward.0.weight: 1048576个参数  encoder.layer.4.feed\_forward.0.bias: 2048个参数  encoder.layer.4.feed\_forward.2.weight: 1048576个参数  encoder.layer.4.feed\_forward.2.bias: 512个参数  encoder.layer.5.attention.values.weight: 4096个参数  encoder.layer.5.attention.keys.weight: 4096个参数  encoder.layer.5.attention.queries.weight: 4096个参数  encoder.layer.5.attention.fc\_out.weight: 262144个参数  encoder.layer.5.attention.fc\_out.bias: 512个参数  encoder.layer.5.norm1.weight: 512个参数  encoder.layer.5.norm1.bias: 512个参数  encoder.layer.5.norm2.weight: 512个参数  encoder.layer.5.norm2.bias: 512个参数  encoder.layer.5.feed\_forward.0.weight: 1048576个参数  encoder.layer.5.feed\_forward.0.bias: 2048个参数  encoder.layer.5.feed\_forward.2.weight: 1048576个参数  encoder.layer.5.feed\_forward.2.bias: 512个参数  decoder.word\_embedding.weight: 3455488个参数  decoder.position\_embedding.weight: 12800个参数  decoder.layers.0.attention.values.weight: 4096个参数  decoder.layers.0.attention.keys.weight: 4096个参数  decoder.layers.0.attention.queries.weight: 4096个参数  decoder.layers.0.attention.fc\_out.weight: 262144个参数  decoder.layers.0.attention.fc\_out.bias: 512个参数  decoder.layers.0.norm.weight: 512个参数  decoder.layers.0.norm.bias: 512个参数  decoder.layers.0.transformer\_block.attention.values.weight: 4096个参数  decoder.layers.0.transformer\_block.attention.keys.weight: 4096个参数  decoder.layers.0.transformer\_block.attention.queries.weight: 4096个参数  decoder.layers.0.transformer\_block.attention.fc\_out.weight: 262144个参数  decoder.layers.0.transformer\_block.attention.fc\_out.bias: 512个参数  decoder.layers.0.transformer\_block.norm1.weight: 512个参数  decoder.layers.0.transformer\_block.norm1.bias: 512个参数  decoder.layers.0.transformer\_block.norm2.weight: 512个参数  decoder.layers.0.transformer\_block.norm2.bias: 512个参数  decoder.layers.0.transformer\_block.feed\_forward.0.weight: 1048576个参数  decoder.layers.0.transformer\_block.feed\_forward.0.bias: 2048个参数  decoder.layers.0.transformer\_block.feed\_forward.2.weight: 1048576个参数  decoder.layers.0.transformer\_block.feed\_forward.2.bias: 512个参数  decoder.layers.1.attention.values.weight: 4096个参数  decoder.layers.1.attention.keys.weight: 4096个参数  decoder.layers.1.attention.queries.weight: 4096个参数  decoder.layers.1.attention.fc\_out.weight: 262144个参数  decoder.layers.1.attention.fc\_out.bias: 512个参数  decoder.layers.1.norm.weight: 512个参数  decoder.layers.1.norm.bias: 512个参数  decoder.layers.1.transformer\_block.attention.values.weight: 4096个参数  decoder.layers.1.transformer\_block.attention.keys.weight: 4096个参数  decoder.layers.1.transformer\_block.attention.queries.weight: 4096个参数  decoder.layers.1.transformer\_block.attention.fc\_out.weight: 262144个参数  decoder.layers.1.transformer\_block.attention.fc\_out.bias: 512个参数  decoder.layers.1.transformer\_block.norm1.weight: 512个参数  decoder.layers.1.transformer\_block.norm1.bias: 512个参数  decoder.layers.1.transformer\_block.norm2.weight: 512个参数  decoder.layers.1.transformer\_block.norm2.bias: 512个参数  decoder.layers.1.transformer\_block.feed\_forward.0.weight: 1048576个参数  decoder.layers.1.transformer\_block.feed\_forward.0.bias: 2048个参数  decoder.layers.1.transformer\_block.feed\_forward.2.weight: 1048576个参数  decoder.layers.1.transformer\_block.feed\_forward.2.bias: 512个参数  decoder.layers.2.attention.values.weight: 4096个参数  decoder.layers.2.attention.keys.weight: 4096个参数  decoder.layers.2.attention.queries.weight: 4096个参数  decoder.layers.2.attention.fc\_out.weight: 262144个参数  decoder.layers.2.attention.fc\_out.bias: 512个参数  decoder.layers.2.norm.weight: 512个参数  decoder.layers.2.norm.bias: 512个参数  decoder.layers.2.transformer\_block.attention.values.weight: 4096个参数  decoder.layers.2.transformer\_block.attention.keys.weight: 4096个参数  decoder.layers.2.transformer\_block.attention.queries.weight: 4096个参数  decoder.layers.2.transformer\_block.attention.fc\_out.weight: 262144个参数  decoder.layers.2.transformer\_block.attention.fc\_out.bias: 512个参数  decoder.layers.2.transformer\_block.norm1.weight: 512个参数  decoder.layers.2.transformer\_block.norm1.bias: 512个参数  decoder.layers.2.transformer\_block.norm2.weight: 512个参数  decoder.layers.2.transformer\_block.norm2.bias: 512个参数  decoder.layers.2.transformer\_block.feed\_forward.0.weight: 1048576个参数  decoder.layers.2.transformer\_block.feed\_forward.0.bias: 2048个参数  decoder.layers.2.transformer\_block.feed\_forward.2.weight: 1048576个参数  decoder.layers.2.transformer\_block.feed\_forward.2.bias: 512个参数  decoder.layers.3.attention.values.weight: 4096个参数  decoder.layers.3.attention.keys.weight: 4096个参数  decoder.layers.3.attention.queries.weight: 4096个参数  decoder.layers.3.attention.fc\_out.weight: 262144个参数  decoder.layers.3.attention.fc\_out.bias: 512个参数  decoder.layers.3.norm.weight: 512个参数  decoder.layers.3.norm.bias: 512个参数  decoder.layers.3.transformer\_block.attention.values.weight: 4096个参数  decoder.layers.3.transformer\_block.attention.keys.weight: 4096个参数  decoder.layers.3.transformer\_block.attention.queries.weight: 4096个参数  decoder.layers.3.transformer\_block.attention.fc\_out.weight: 262144个参数  decoder.layers.3.transformer\_block.attention.fc\_out.bias: 512个参数  decoder.layers.3.transformer\_block.norm1.weight: 512个参数  decoder.layers.3.transformer\_block.norm1.bias: 512个参数  decoder.layers.3.transformer\_block.norm2.weight: 512个参数  decoder.layers.3.transformer\_block.norm2.bias: 512个参数  decoder.layers.3.transformer\_block.feed\_forward.0.weight: 1048576个参数  decoder.layers.3.transformer\_block.feed\_forward.0.bias: 2048个参数  decoder.layers.3.transformer\_block.feed\_forward.2.weight: 1048576个参数  decoder.layers.3.transformer\_block.feed\_forward.2.bias: 512个参数  decoder.layers.4.attention.values.weight: 4096个参数  decoder.layers.4.attention.keys.weight: 4096个参数  decoder.layers.4.attention.queries.weight: 4096个参数  decoder.layers.4.attention.fc\_out.weight: 262144个参数  decoder.layers.4.attention.fc\_out.bias: 512个参数  decoder.layers.4.norm.weight: 512个参数  decoder.layers.4.norm.bias: 512个参数  decoder.layers.4.transformer\_block.attention.values.weight: 4096个参数  decoder.layers.4.transformer\_block.attention.keys.weight: 4096个参数  decoder.layers.4.transformer\_block.attention.queries.weight: 4096个参数  decoder.layers.4.transformer\_block.attention.fc\_out.weight: 262144个参数  decoder.layers.4.transformer\_block.attention.fc\_out.bias: 512个参数  decoder.layers.4.transformer\_block.norm1.weight: 512个参数  decoder.layers.4.transformer\_block.norm1.bias: 512个参数  decoder.layers.4.transformer\_block.norm2.weight: 512个参数  decoder.layers.4.transformer\_block.norm2.bias: 512个参数  decoder.layers.4.transformer\_block.feed\_forward.0.weight: 1048576个参数  decoder.layers.4.transformer\_block.feed\_forward.0.bias: 2048个参数  decoder.layers.4.transformer\_block.feed\_forward.2.weight: 1048576个参数  decoder.layers.4.transformer\_block.feed\_forward.2.bias: 512个参数  decoder.layers.5.attention.values.weight: 4096个参数  decoder.layers.5.attention.keys.weight: 4096个参数  decoder.layers.5.attention.queries.weight: 4096个参数  decoder.layers.5.attention.fc\_out.weight: 262144个参数  decoder.layers.5.attention.fc\_out.bias: 512个参数  decoder.layers.5.norm.weight: 512个参数  decoder.layers.5.norm.bias: 512个参数  decoder.layers.5.transformer\_block.attention.values.weight: 4096个参数  decoder.layers.5.transformer\_block.attention.keys.weight: 4096个参数  decoder.layers.5.transformer\_block.attention.queries.weight: 4096个参数  decoder.layers.5.transformer\_block.attention.fc\_out.weight: 262144个参数  decoder.layers.5.transformer\_block.attention.fc\_out.bias: 512个参数  decoder.layers.5.transformer\_block.norm1.weight: 512个参数  decoder.layers.5.transformer\_block.norm1.bias: 512个参数  decoder.layers.5.transformer\_block.norm2.weight: 512个参数  decoder.layers.5.transformer\_block.norm2.bias: 512个参数  decoder.layers.5.transformer\_block.feed\_forward.0.weight: 1048576个参数  decoder.layers.5.transformer\_block.feed\_forward.0.bias: 2048个参数  decoder.layers.5.transformer\_block.feed\_forward.2.weight: 1048576个参数  decoder.layers.5.transformer\_block.feed\_forward.2.bias: 512个参数  decoder.fc\_out.weight: 3455488个参数  decoder.fc\_out.bias: 6749个参数  总参数数量: 42263133个 |

**五、trochinfo输出模型信息**

1.方法：

torchinfo本身并不依赖特定的配置文件运行。其配置选项多通过函数参数传递，在调用summary()时可根据需要调整其行为。例如，改变模型摘要输出的详细程度：

使用代码可以如下：

|  |
| --- |
| summary(model, verbose=2) # 更详细的输出 |

注意：这里verbose参数决定输出信息的详尽度，不同的数值代表不同的信息深度。verbose must be either 0 (quiet), 1 (default), or 2 (verbose).

2.实际调用

① 代码与输出结果如下：

|  |
| --- |
| summary(model) |

|  |
| --- |
| 模型参数信息：  ===========================================================================  Layer (type:depth-idx) Param #  ===========================================================================  Transformer --  ├─Encoder: 1-1 --  │ └─Embedding: 2-1 5,143,552  │ └─Embedding: 2-2 12,800  │ └─ModuleList: 2-3 --  │ │ └─TransformerBlock: 3-1 2,376,704  │ │ └─TransformerBlock: 3-2 2,376,704  │ │ └─TransformerBlock: 3-3 2,376,704  │ │ └─TransformerBlock: 3-4 2,376,704  │ │ └─TransformerBlock: 3-5 2,376,704  │ │ └─TransformerBlock: 3-6 2,376,704  │ └─Dropout: 2-4 --  ├─Decoder: 1-2 --  │ └─Embedding: 2-5 3,455,488  │ └─Embedding: 2-6 12,800  │ └─ModuleList: 2-7 --  │ │ └─DecoderBlock: 3-7 2,652,672  │ │ └─DecoderBlock: 3-8 2,652,672  │ │ └─DecoderBlock: 3-9 2,652,672  │ │ └─DecoderBlock: 3-10 2,652,672  │ │ └─DecoderBlock: 3-11 2,652,672  │ │ └─DecoderBlock: 3-12 2,652,672  │ └─Linear: 2-8 3,462,237  │ └─Dropout: 2-9 --  ===========================================================================  Total params: 42,263,133  Trainable params: 42,263,133  Non-trainable params: 0  =========================================================================== |

②更详细的结果；

|  |
| --- |
| summary(model, *verbose*=2) |

|  |
| --- |
| 模型参数信息：  ===========================================================================  Layer (type:depth-idx) Param #  ===========================================================================  Transformer --  ├─Encoder: 1-1 --  │ └─word\_embedding.weight ├─5,143,552  │ └─position\_embedding.weight ├─12,800  │ └─layer.0.attention.values.weight ├─4,096  │ └─layer.0.attention.keys.weight ├─4,096  │ └─layer.0.attention.queries.weight ├─4,096  │ └─layer.0.attention.fc\_out.weight ├─262,144  │ └─layer.0.attention.fc\_out.bias ├─512  │ └─layer.0.norm1.weight ├─512  │ └─layer.0.norm1.bias ├─512  │ └─layer.0.norm2.weight ├─512  │ └─layer.0.norm2.bias ├─512  │ └─layer.0.feed\_forward.0.weight ├─1,048,576  │ └─layer.0.feed\_forward.0.bias ├─2,048  │ └─layer.0.feed\_forward.2.weight ├─1,048,576  │ └─layer.0.feed\_forward.2.bias ├─512  │ └─layer.1.attention.values.weight ├─4,096  │ └─layer.1.attention.keys.weight ├─4,096  │ └─layer.1.attention.queries.weight ├─4,096  │ └─layer.1.attention.fc\_out.weight ├─262,144  │ └─layer.1.attention.fc\_out.bias ├─512  │ └─layer.1.norm1.weight ├─512  │ └─layer.1.norm1.bias ├─512  │ └─layer.1.norm2.weight ├─512  │ └─layer.1.norm2.bias ├─512  │ └─layer.1.feed\_forward.0.weight ├─1,048,576  │ └─layer.1.feed\_forward.0.bias ├─2,048  │ └─layer.1.feed\_forward.2.weight ├─1,048,576  │ └─layer.1.feed\_forward.2.bias ├─512  │ └─layer.2.attention.values.weight ├─4,096  │ └─layer.2.attention.keys.weight ├─4,096  │ └─layer.2.attention.queries.weight ├─4,096  │ └─layer.2.attention.fc\_out.weight ├─262,144  │ └─layer.2.attention.fc\_out.bias ├─512  │ └─layer.2.norm1.weight ├─512  │ └─layer.2.norm1.bias ├─512  │ └─layer.2.norm2.weight ├─512  │ └─layer.2.norm2.bias ├─512  │ └─layer.2.feed\_forward.0.weight ├─1,048,576  │ └─layer.2.feed\_forward.0.bias ├─2,048  │ └─layer.2.feed\_forward.2.weight ├─1,048,576  │ └─layer.2.feed\_forward.2.bias ├─512  │ └─layer.3.attention.values.weight ├─4,096  │ └─layer.3.attention.keys.weight ├─4,096  │ └─layer.3.attention.queries.weight ├─4,096  │ └─layer.3.attention.fc\_out.weight ├─262,144  │ └─layer.3.attention.fc\_out.bias ├─512  │ └─layer.3.norm1.weight ├─512  │ └─layer.3.norm1.bias ├─512  │ └─layer.3.norm2.weight ├─512  │ └─layer.3.norm2.bias ├─512  │ └─layer.3.feed\_forward.0.weight ├─1,048,576  │ └─layer.3.feed\_forward.0.bias ├─2,048  │ └─layer.3.feed\_forward.2.weight ├─1,048,576  │ └─layer.3.feed\_forward.2.bias ├─512  │ └─layer.4.attention.values.weight ├─4,096  │ └─layer.4.attention.keys.weight ├─4,096  │ └─layer.4.attention.queries.weight ├─4,096  │ └─layer.4.attention.fc\_out.weight ├─262,144  │ └─layer.4.attention.fc\_out.bias ├─512  │ └─layer.4.norm1.weight ├─512  │ └─layer.4.norm1.bias ├─512  │ └─layer.4.norm2.weight ├─512  │ └─layer.4.norm2.bias ├─512  │ └─layer.4.feed\_forward.0.weight ├─1,048,576  │ └─layer.4.feed\_forward.0.bias ├─2,048  │ └─layer.4.feed\_forward.2.weight ├─1,048,576  │ └─layer.4.feed\_forward.2.bias ├─512  │ └─layer.5.attention.values.weight ├─4,096  │ └─layer.5.attention.keys.weight ├─4,096  │ └─layer.5.attention.queries.weight ├─4,096  │ └─layer.5.attention.fc\_out.weight ├─262,144  │ └─layer.5.attention.fc\_out.bias ├─512  │ └─layer.5.norm1.weight ├─512  │ └─layer.5.norm1.bias ├─512  │ └─layer.5.norm2.weight ├─512  │ └─layer.5.norm2.bias ├─512  │ └─layer.5.feed\_forward.0.weight ├─1,048,576  │ └─layer.5.feed\_forward.0.bias ├─2,048  │ └─layer.5.feed\_forward.2.weight ├─1,048,576  │ └─layer.5.feed\_forward.2.bias └─512  │ └─Embedding: 2-1 5,143,552  │ │ └─weight └─5,143,552  │ └─Embedding: 2-2 12,800  │ │ └─weight └─12,800  │ └─ModuleList: 2-3 --  │ │ └─0.attention.values.weight ├─4,096  │ │ └─0.attention.keys.weight ├─4,096  │ │ └─0.attention.queries.weight ├─4,096  │ │ └─0.attention.fc\_out.weight ├─262,144  │ │ └─0.attention.fc\_out.bias ├─512  │ │ └─0.norm1.weight ├─512  │ │ └─0.norm1.bias ├─512  │ │ └─0.norm2.weight ├─512  │ │ └─0.norm2.bias ├─512  │ │ └─0.feed\_forward.0.weight ├─1,048,576  │ │ └─0.feed\_forward.0.bias ├─2,048  │ │ └─0.feed\_forward.2.weight ├─1,048,576  │ │ └─0.feed\_forward.2.bias ├─512  │ │ └─1.attention.values.weight ├─4,096  │ │ └─1.attention.keys.weight ├─4,096  │ │ └─1.attention.queries.weight ├─4,096  │ │ └─1.attention.fc\_out.weight ├─262,144  │ │ └─1.attention.fc\_out.bias ├─512  │ │ └─1.norm1.weight ├─512  │ │ └─1.norm1.bias ├─512  │ │ └─1.norm2.weight ├─512  │ │ └─1.norm2.bias ├─512  │ │ └─1.feed\_forward.0.weight ├─1,048,576  │ │ └─1.feed\_forward.0.bias ├─2,048  │ │ └─1.feed\_forward.2.weight ├─1,048,576  │ │ └─1.feed\_forward.2.bias ├─512  │ │ └─2.attention.values.weight ├─4,096  │ │ └─2.attention.keys.weight ├─4,096  │ │ └─2.attention.queries.weight ├─4,096  │ │ └─2.attention.fc\_out.weight ├─262,144  │ │ └─2.attention.fc\_out.bias ├─512  │ │ └─2.norm1.weight ├─512  │ │ └─2.norm1.bias ├─512  │ │ └─2.norm2.weight ├─512  │ │ └─2.norm2.bias ├─512  │ │ └─2.feed\_forward.0.weight ├─1,048,576  │ │ └─2.feed\_forward.0.bias ├─2,048  │ │ └─2.feed\_forward.2.weight ├─1,048,576  │ │ └─2.feed\_forward.2.bias ├─512  │ │ └─3.attention.values.weight ├─4,096  │ │ └─3.attention.keys.weight ├─4,096  │ │ └─3.attention.queries.weight ├─4,096  │ │ └─3.attention.fc\_out.weight ├─262,144  │ │ └─3.attention.fc\_out.bias ├─512  │ │ └─3.norm1.weight ├─512  │ │ └─3.norm1.bias ├─512  │ │ └─3.norm2.weight ├─512  │ │ └─3.norm2.bias ├─512  │ │ └─3.feed\_forward.0.weight ├─1,048,576  │ │ └─3.feed\_forward.0.bias ├─2,048  │ │ └─3.feed\_forward.2.weight ├─1,048,576  │ │ └─3.feed\_forward.2.bias ├─512  │ │ └─4.attention.values.weight ├─4,096  │ │ └─4.attention.keys.weight ├─4,096  │ │ └─4.attention.queries.weight ├─4,096  │ │ └─4.attention.fc\_out.weight ├─262,144  │ │ └─4.attention.fc\_out.bias ├─512  │ │ └─4.norm1.weight ├─512  │ │ └─4.norm1.bias ├─512  │ │ └─4.norm2.weight ├─512  │ │ └─4.norm2.bias ├─512  │ │ └─4.feed\_forward.0.weight ├─1,048,576  │ │ └─4.feed\_forward.0.bias ├─2,048  │ │ └─4.feed\_forward.2.weight ├─1,048,576  │ │ └─4.feed\_forward.2.bias ├─512  │ │ └─5.attention.values.weight ├─4,096  │ │ └─5.attention.keys.weight ├─4,096  │ │ └─5.attention.queries.weight ├─4,096  │ │ └─5.attention.fc\_out.weight ├─262,144  │ │ └─5.attention.fc\_out.bias ├─512  │ │ └─5.norm1.weight ├─512  │ │ └─5.norm1.bias ├─512  │ │ └─5.norm2.weight ├─512  │ │ └─5.norm2.bias ├─512  │ │ └─5.feed\_forward.0.weight ├─1,048,576  │ │ └─5.feed\_forward.0.bias ├─2,048  │ │ └─5.feed\_forward.2.weight ├─1,048,576  │ │ └─5.feed\_forward.2.bias └─512  │ │ └─TransformerBlock: 3-1 2,376,704  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm1.weight ├─512  │ │ │ └─norm1.bias ├─512  │ │ │ └─norm2.weight ├─512  │ │ │ └─norm2.bias ├─512  │ │ │ └─feed\_forward.0.weight ├─1,048,576  │ │ │ └─feed\_forward.0.bias ├─2,048  │ │ │ └─feed\_forward.2.weight ├─1,048,576  │ │ │ └─feed\_forward.2.bias └─512  │ │ └─TransformerBlock: 3-2 2,376,704  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm1.weight ├─512  │ │ │ └─norm1.bias ├─512  │ │ │ └─norm2.weight ├─512  │ │ │ └─norm2.bias ├─512  │ │ │ └─feed\_forward.0.weight ├─1,048,576  │ │ │ └─feed\_forward.0.bias ├─2,048  │ │ │ └─feed\_forward.2.weight ├─1,048,576  │ │ │ └─feed\_forward.2.bias └─512  │ │ └─TransformerBlock: 3-3 2,376,704  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm1.weight ├─512  │ │ │ └─norm1.bias ├─512  │ │ │ └─norm2.weight ├─512  │ │ │ └─norm2.bias ├─512  │ │ │ └─feed\_forward.0.weight ├─1,048,576  │ │ │ └─feed\_forward.0.bias ├─2,048  │ │ │ └─feed\_forward.2.weight ├─1,048,576  │ │ │ └─feed\_forward.2.bias └─512  │ │ └─TransformerBlock: 3-4 2,376,704  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm1.weight ├─512  │ │ │ └─norm1.bias ├─512  │ │ │ └─norm2.weight ├─512  │ │ │ └─norm2.bias ├─512  │ │ │ └─feed\_forward.0.weight ├─1,048,576  │ │ │ └─feed\_forward.0.bias ├─2,048  │ │ │ └─feed\_forward.2.weight ├─1,048,576  │ │ │ └─feed\_forward.2.bias └─512  │ │ └─TransformerBlock: 3-5 2,376,704  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm1.weight ├─512  │ │ │ └─norm1.bias ├─512  │ │ │ └─norm2.weight ├─512  │ │ │ └─norm2.bias ├─512  │ │ │ └─feed\_forward.0.weight ├─1,048,576  │ │ │ └─feed\_forward.0.bias ├─2,048  │ │ │ └─feed\_forward.2.weight ├─1,048,576  │ │ │ └─feed\_forward.2.bias └─512  │ │ └─TransformerBlock: 3-6 2,376,704  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm1.weight ├─512  │ │ │ └─norm1.bias ├─512  │ │ │ └─norm2.weight ├─512  │ │ │ └─norm2.bias ├─512  │ │ │ └─feed\_forward.0.weight ├─1,048,576  │ │ │ └─feed\_forward.0.bias ├─2,048  │ │ │ └─feed\_forward.2.weight ├─1,048,576  │ │ │ └─feed\_forward.2.bias └─512  │ └─Dropout: 2-4 --  ├─Decoder: 1-2 --  │ └─word\_embedding.weight ├─3,455,488  │ └─position\_embedding.weight ├─12,800  │ └─layers.0.attention.values.weight ├─4,096  │ └─layers.0.attention.keys.weight ├─4,096  │ └─layers.0.attention.queries.weight ├─4,096  │ └─layers.0.attention.fc\_out.weight ├─262,144  │ └─layers.0.attention.fc\_out.bias ├─512  │ └─layers.0.norm.weight ├─512  │ └─layers.0.norm.bias ├─512  │ └─layers.0.transformer\_block.attention.values.weight ├─4,096  │ └─layers.0.transformer\_block.attention.keys.weight ├─4,096  │ └─layers.0.transformer\_block.attention.queries.weight ├─4,096  │ └─layers.0.transformer\_block.attention.fc\_out.weight ├─262,144  │ └─layers.0.transformer\_block.attention.fc\_out.bias ├─512  │ └─layers.0.transformer\_block.norm1.weight ├─512  │ └─layers.0.transformer\_block.norm1.bias ├─512  │ └─layers.0.transformer\_block.norm2.weight ├─512  │ └─layers.0.transformer\_block.norm2.bias ├─512  │ └─layers.0.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ └─layers.0.transformer\_block.feed\_forward.0.bias ├─2,048  │ └─layers.0.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ └─layers.0.transformer\_block.feed\_forward.2.bias ├─512  │ └─layers.1.attention.values.weight ├─4,096  │ └─layers.1.attention.keys.weight ├─4,096  │ └─layers.1.attention.queries.weight ├─4,096  │ └─layers.1.attention.fc\_out.weight ├─262,144  │ └─layers.1.attention.fc\_out.bias ├─512  │ └─layers.1.norm.weight ├─512  │ └─layers.1.norm.bias ├─512  │ └─layers.1.transformer\_block.attention.values.weight ├─4,096  │ └─layers.1.transformer\_block.attention.keys.weight ├─4,096  │ └─layers.1.transformer\_block.attention.queries.weight ├─4,096  │ └─layers.1.transformer\_block.attention.fc\_out.weight ├─262,144  │ └─layers.1.transformer\_block.attention.fc\_out.bias ├─512  │ └─layers.1.transformer\_block.norm1.weight ├─512  │ └─layers.1.transformer\_block.norm1.bias ├─512  │ └─layers.1.transformer\_block.norm2.weight ├─512  │ └─layers.1.transformer\_block.norm2.bias ├─512  │ └─layers.1.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ └─layers.1.transformer\_block.feed\_forward.0.bias ├─2,048  │ └─layers.1.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ └─layers.1.transformer\_block.feed\_forward.2.bias ├─512  │ └─layers.2.attention.values.weight ├─4,096  │ └─layers.2.attention.keys.weight ├─4,096  │ └─layers.2.attention.queries.weight ├─4,096  │ └─layers.2.attention.fc\_out.weight ├─262,144  │ └─layers.2.attention.fc\_out.bias ├─512  │ └─layers.2.norm.weight ├─512  │ └─layers.2.norm.bias ├─512  │ └─layers.2.transformer\_block.attention.values.weight ├─4,096  │ └─layers.2.transformer\_block.attention.keys.weight ├─4,096  │ └─layers.2.transformer\_block.attention.queries.weight ├─4,096  │ └─layers.2.transformer\_block.attention.fc\_out.weight ├─262,144  │ └─layers.2.transformer\_block.attention.fc\_out.bias ├─512  │ └─layers.2.transformer\_block.norm1.weight ├─512  │ └─layers.2.transformer\_block.norm1.bias ├─512  │ └─layers.2.transformer\_block.norm2.weight ├─512  │ └─layers.2.transformer\_block.norm2.bias ├─512  │ └─layers.2.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ └─layers.2.transformer\_block.feed\_forward.0.bias ├─2,048  │ └─layers.2.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ └─layers.2.transformer\_block.feed\_forward.2.bias ├─512  │ └─layers.3.attention.values.weight ├─4,096  │ └─layers.3.attention.keys.weight ├─4,096  │ └─layers.3.attention.queries.weight ├─4,096  │ └─layers.3.attention.fc\_out.weight ├─262,144  │ └─layers.3.attention.fc\_out.bias ├─512  │ └─layers.3.norm.weight ├─512  │ └─layers.3.norm.bias ├─512  │ └─layers.3.transformer\_block.attention.values.weight ├─4,096  │ └─layers.3.transformer\_block.attention.keys.weight ├─4,096  │ └─layers.3.transformer\_block.attention.queries.weight ├─4,096  │ └─layers.3.transformer\_block.attention.fc\_out.weight ├─262,144  │ └─layers.3.transformer\_block.attention.fc\_out.bias ├─512  │ └─layers.3.transformer\_block.norm1.weight ├─512  │ └─layers.3.transformer\_block.norm1.bias ├─512  │ └─layers.3.transformer\_block.norm2.weight ├─512  │ └─layers.3.transformer\_block.norm2.bias ├─512  │ └─layers.3.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ └─layers.3.transformer\_block.feed\_forward.0.bias ├─2,048  │ └─layers.3.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ └─layers.3.transformer\_block.feed\_forward.2.bias ├─512  │ └─layers.4.attention.values.weight ├─4,096  │ └─layers.4.attention.keys.weight ├─4,096  │ └─layers.4.attention.queries.weight ├─4,096  │ └─layers.4.attention.fc\_out.weight ├─262,144  │ └─layers.4.attention.fc\_out.bias ├─512  │ └─layers.4.norm.weight ├─512  │ └─layers.4.norm.bias ├─512  │ └─layers.4.transformer\_block.attention.values.weight ├─4,096  │ └─layers.4.transformer\_block.attention.keys.weight ├─4,096  │ └─layers.4.transformer\_block.attention.queries.weight ├─4,096  │ └─layers.4.transformer\_block.attention.fc\_out.weight ├─262,144  │ └─layers.4.transformer\_block.attention.fc\_out.bias ├─512  │ └─layers.4.transformer\_block.norm1.weight ├─512  │ └─layers.4.transformer\_block.norm1.bias ├─512  │ └─layers.4.transformer\_block.norm2.weight ├─512  │ └─layers.4.transformer\_block.norm2.bias ├─512  │ └─layers.4.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ └─layers.4.transformer\_block.feed\_forward.0.bias ├─2,048  │ └─layers.4.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ └─layers.4.transformer\_block.feed\_forward.2.bias ├─512  │ └─layers.5.attention.values.weight ├─4,096  │ └─layers.5.attention.keys.weight ├─4,096  │ └─layers.5.attention.queries.weight ├─4,096  │ └─layers.5.attention.fc\_out.weight ├─262,144  │ └─layers.5.attention.fc\_out.bias ├─512  │ └─layers.5.norm.weight ├─512  │ └─layers.5.norm.bias ├─512  │ └─layers.5.transformer\_block.attention.values.weight ├─4,096  │ └─layers.5.transformer\_block.attention.keys.weight ├─4,096  │ └─layers.5.transformer\_block.attention.queries.weight ├─4,096  │ └─layers.5.transformer\_block.attention.fc\_out.weight ├─262,144  │ └─layers.5.transformer\_block.attention.fc\_out.bias ├─512  │ └─layers.5.transformer\_block.norm1.weight ├─512  │ └─layers.5.transformer\_block.norm1.bias ├─512  │ └─layers.5.transformer\_block.norm2.weight ├─512  │ └─layers.5.transformer\_block.norm2.bias ├─512  │ └─layers.5.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ └─layers.5.transformer\_block.feed\_forward.0.bias ├─2,048  │ └─layers.5.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ └─layers.5.transformer\_block.feed\_forward.2.bias ├─512  │ └─fc\_out.weight ├─3,455,488  │ └─fc\_out.bias └─6,749  │ └─Embedding: 2-5 3,455,488  │ │ └─weight └─3,455,488  │ └─Embedding: 2-6 12,800  │ │ └─weight └─12,800  │ └─ModuleList: 2-7 --  │ │ └─0.attention.values.weight ├─4,096  │ │ └─0.attention.keys.weight ├─4,096  │ │ └─0.attention.queries.weight ├─4,096  │ │ └─0.attention.fc\_out.weight ├─262,144  │ │ └─0.attention.fc\_out.bias ├─512  │ │ └─0.norm.weight ├─512  │ │ └─0.norm.bias ├─512  │ │ └─0.transformer\_block.attention.values.weight ├─4,096  │ │ └─0.transformer\_block.attention.keys.weight ├─4,096  │ │ └─0.transformer\_block.attention.queries.weight ├─4,096  │ │ └─0.transformer\_block.attention.fc\_out.weight ├─262,144  │ │ └─0.transformer\_block.attention.fc\_out.bias ├─512  │ │ └─0.transformer\_block.norm1.weight ├─512  │ │ └─0.transformer\_block.norm1.bias ├─512  │ │ └─0.transformer\_block.norm2.weight ├─512  │ │ └─0.transformer\_block.norm2.bias ├─512  │ │ └─0.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ └─0.transformer\_block.feed\_forward.0.bias ├─2,048  │ │ └─0.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ └─0.transformer\_block.feed\_forward.2.bias ├─512  │ │ └─1.attention.values.weight ├─4,096  │ │ └─1.attention.keys.weight ├─4,096  │ │ └─1.attention.queries.weight ├─4,096  │ │ └─1.attention.fc\_out.weight ├─262,144  │ │ └─1.attention.fc\_out.bias ├─512  │ │ └─1.norm.weight ├─512  │ │ └─1.norm.bias ├─512  │ │ └─1.transformer\_block.attention.values.weight ├─4,096  │ │ └─1.transformer\_block.attention.keys.weight ├─4,096  │ │ └─1.transformer\_block.attention.queries.weight ├─4,096  │ │ └─1.transformer\_block.attention.fc\_out.weight ├─262,144  │ │ └─1.transformer\_block.attention.fc\_out.bias ├─512  │ │ └─1.transformer\_block.norm1.weight ├─512  │ │ └─1.transformer\_block.norm1.bias ├─512  │ │ └─1.transformer\_block.norm2.weight ├─512  │ │ └─1.transformer\_block.norm2.bias ├─512  │ │ └─1.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ └─1.transformer\_block.feed\_forward.0.bias ├─2,048  │ │ └─1.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ └─1.transformer\_block.feed\_forward.2.bias ├─512  │ │ └─2.attention.values.weight ├─4,096  │ │ └─2.attention.keys.weight ├─4,096  │ │ └─2.attention.queries.weight ├─4,096  │ │ └─2.attention.fc\_out.weight ├─262,144  │ │ └─2.attention.fc\_out.bias ├─512  │ │ └─2.norm.weight ├─512  │ │ └─2.norm.bias ├─512  │ │ └─2.transformer\_block.attention.values.weight ├─4,096  │ │ └─2.transformer\_block.attention.keys.weight ├─4,096  │ │ └─2.transformer\_block.attention.queries.weight ├─4,096  │ │ └─2.transformer\_block.attention.fc\_out.weight ├─262,144  │ │ └─2.transformer\_block.attention.fc\_out.bias ├─512  │ │ └─2.transformer\_block.norm1.weight ├─512  │ │ └─2.transformer\_block.norm1.bias ├─512  │ │ └─2.transformer\_block.norm2.weight ├─512  │ │ └─2.transformer\_block.norm2.bias ├─512  │ │ └─2.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ └─2.transformer\_block.feed\_forward.0.bias ├─2,048  │ │ └─2.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ └─2.transformer\_block.feed\_forward.2.bias ├─512  │ │ └─3.attention.values.weight ├─4,096  │ │ └─3.attention.keys.weight ├─4,096  │ │ └─3.attention.queries.weight ├─4,096  │ │ └─3.attention.fc\_out.weight ├─262,144  │ │ └─3.attention.fc\_out.bias ├─512  │ │ └─3.norm.weight ├─512  │ │ └─3.norm.bias ├─512  │ │ └─3.transformer\_block.attention.values.weight ├─4,096  │ │ └─3.transformer\_block.attention.keys.weight ├─4,096  │ │ └─3.transformer\_block.attention.queries.weight ├─4,096  │ │ └─3.transformer\_block.attention.fc\_out.weight ├─262,144  │ │ └─3.transformer\_block.attention.fc\_out.bias ├─512  │ │ └─3.transformer\_block.norm1.weight ├─512  │ │ └─3.transformer\_block.norm1.bias ├─512  │ │ └─3.transformer\_block.norm2.weight ├─512  │ │ └─3.transformer\_block.norm2.bias ├─512  │ │ └─3.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ └─3.transformer\_block.feed\_forward.0.bias ├─2,048  │ │ └─3.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ └─3.transformer\_block.feed\_forward.2.bias ├─512  │ │ └─4.attention.values.weight ├─4,096  │ │ └─4.attention.keys.weight ├─4,096  │ │ └─4.attention.queries.weight ├─4,096  │ │ └─4.attention.fc\_out.weight ├─262,144  │ │ └─4.attention.fc\_out.bias ├─512  │ │ └─4.norm.weight ├─512  │ │ └─4.norm.bias ├─512  │ │ └─4.transformer\_block.attention.values.weight ├─4,096  │ │ └─4.transformer\_block.attention.keys.weight ├─4,096  │ │ └─4.transformer\_block.attention.queries.weight ├─4,096  │ │ └─4.transformer\_block.attention.fc\_out.weight ├─262,144  │ │ └─4.transformer\_block.attention.fc\_out.bias ├─512  │ │ └─4.transformer\_block.norm1.weight ├─512  │ │ └─4.transformer\_block.norm1.bias ├─512  │ │ └─4.transformer\_block.norm2.weight ├─512  │ │ └─4.transformer\_block.norm2.bias ├─512  │ │ └─4.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ └─4.transformer\_block.feed\_forward.0.bias ├─2,048  │ │ └─4.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ └─4.transformer\_block.feed\_forward.2.bias ├─512  │ │ └─5.attention.values.weight ├─4,096  │ │ └─5.attention.keys.weight ├─4,096  │ │ └─5.attention.queries.weight ├─4,096  │ │ └─5.attention.fc\_out.weight ├─262,144  │ │ └─5.attention.fc\_out.bias ├─512  │ │ └─5.norm.weight ├─512  │ │ └─5.norm.bias ├─512  │ │ └─5.transformer\_block.attention.values.weight ├─4,096  │ │ └─5.transformer\_block.attention.keys.weight ├─4,096  │ │ └─5.transformer\_block.attention.queries.weight ├─4,096  │ │ └─5.transformer\_block.attention.fc\_out.weight ├─262,144  │ │ └─5.transformer\_block.attention.fc\_out.bias ├─512  │ │ └─5.transformer\_block.norm1.weight ├─512  │ │ └─5.transformer\_block.norm1.bias ├─512  │ │ └─5.transformer\_block.norm2.weight ├─512  │ │ └─5.transformer\_block.norm2.bias ├─512  │ │ └─5.transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ └─5.transformer\_block.feed\_forward.0.bias ├─2,048  │ │ └─5.transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ └─5.transformer\_block.feed\_forward.2.bias └─512  │ │ └─DecoderBlock: 3-7 2,652,672  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm.weight ├─512  │ │ │ └─norm.bias ├─512  │ │ │ └─transformer\_block.attention.values.weight ├─4,096  │ │ │ └─transformer\_block.attention.keys.weight ├─4,096  │ │ │ └─transformer\_block.attention.queries.weight ├─4,096  │ │ │ └─transformer\_block.attention.fc\_out.weight ├─262,144  │ │ │ └─transformer\_block.attention.fc\_out.bias ├─512  │ │ │ └─transformer\_block.norm1.weight ├─512  │ │ │ └─transformer\_block.norm1.bias ├─512  │ │ │ └─transformer\_block.norm2.weight ├─512  │ │ │ └─transformer\_block.norm2.bias ├─512  │ │ │ └─transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.0.bias ├─2,048  │ │ │ └─transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.2.bias └─512  │ │ └─DecoderBlock: 3-8 2,652,672  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm.weight ├─512  │ │ │ └─norm.bias ├─512  │ │ │ └─transformer\_block.attention.values.weight ├─4,096  │ │ │ └─transformer\_block.attention.keys.weight ├─4,096  │ │ │ └─transformer\_block.attention.queries.weight ├─4,096  │ │ │ └─transformer\_block.attention.fc\_out.weight ├─262,144  │ │ │ └─transformer\_block.attention.fc\_out.bias ├─512  │ │ │ └─transformer\_block.norm1.weight ├─512  │ │ │ └─transformer\_block.norm1.bias ├─512  │ │ │ └─transformer\_block.norm2.weight ├─512  │ │ │ └─transformer\_block.norm2.bias ├─512  │ │ │ └─transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.0.bias ├─2,048  │ │ │ └─transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.2.bias └─512  │ │ └─DecoderBlock: 3-9 2,652,672  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm.weight ├─512  │ │ │ └─norm.bias ├─512  │ │ │ └─transformer\_block.attention.values.weight ├─4,096  │ │ │ └─transformer\_block.attention.keys.weight ├─4,096  │ │ │ └─transformer\_block.attention.queries.weight ├─4,096  │ │ │ └─transformer\_block.attention.fc\_out.weight ├─262,144  │ │ │ └─transformer\_block.attention.fc\_out.bias ├─512  │ │ │ └─transformer\_block.norm1.weight ├─512  │ │ │ └─transformer\_block.norm1.bias ├─512  │ │ │ └─transformer\_block.norm2.weight ├─512  │ │ │ └─transformer\_block.norm2.bias ├─512  │ │ │ └─transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.0.bias ├─2,048  │ │ │ └─transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.2.bias └─512  │ │ └─DecoderBlock: 3-10 2,652,672  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm.weight ├─512  │ │ │ └─norm.bias ├─512  │ │ │ └─transformer\_block.attention.values.weight ├─4,096  │ │ │ └─transformer\_block.attention.keys.weight ├─4,096  │ │ │ └─transformer\_block.attention.queries.weight ├─4,096  │ │ │ └─transformer\_block.attention.fc\_out.weight ├─262,144  │ │ │ └─transformer\_block.attention.fc\_out.bias ├─512  │ │ │ └─transformer\_block.norm1.weight ├─512  │ │ │ └─transformer\_block.norm1.bias ├─512  │ │ │ └─transformer\_block.norm2.weight ├─512  │ │ │ └─transformer\_block.norm2.bias ├─512  │ │ │ └─transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.0.bias ├─2,048  │ │ │ └─transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.2.bias └─512  │ │ └─DecoderBlock: 3-11 2,652,672  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm.weight ├─512  │ │ │ └─norm.bias ├─512  │ │ │ └─transformer\_block.attention.values.weight ├─4,096  │ │ │ └─transformer\_block.attention.keys.weight ├─4,096  │ │ │ └─transformer\_block.attention.queries.weight ├─4,096  │ │ │ └─transformer\_block.attention.fc\_out.weight ├─262,144  │ │ │ └─transformer\_block.attention.fc\_out.bias ├─512  │ │ │ └─transformer\_block.norm1.weight ├─512  │ │ │ └─transformer\_block.norm1.bias ├─512  │ │ │ └─transformer\_block.norm2.weight ├─512  │ │ │ └─transformer\_block.norm2.bias ├─512  │ │ │ └─transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.0.bias ├─2,048  │ │ │ └─transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.2.bias └─512  │ │ └─DecoderBlock: 3-12 2,652,672  │ │ │ └─attention.values.weight ├─4,096  │ │ │ └─attention.keys.weight ├─4,096  │ │ │ └─attention.queries.weight ├─4,096  │ │ │ └─attention.fc\_out.weight ├─262,144  │ │ │ └─attention.fc\_out.bias ├─512  │ │ │ └─norm.weight ├─512  │ │ │ └─norm.bias ├─512  │ │ │ └─transformer\_block.attention.values.weight ├─4,096  │ │ │ └─transformer\_block.attention.keys.weight ├─4,096  │ │ │ └─transformer\_block.attention.queries.weight ├─4,096  │ │ │ └─transformer\_block.attention.fc\_out.weight ├─262,144  │ │ │ └─transformer\_block.attention.fc\_out.bias ├─512  │ │ │ └─transformer\_block.norm1.weight ├─512  │ │ │ └─transformer\_block.norm1.bias ├─512  │ │ │ └─transformer\_block.norm2.weight ├─512  │ │ │ └─transformer\_block.norm2.bias ├─512  │ │ │ └─transformer\_block.feed\_forward.0.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.0.bias ├─2,048  │ │ │ └─transformer\_block.feed\_forward.2.weight ├─1,048,576  │ │ │ └─transformer\_block.feed\_forward.2.bias └─512  │ └─Linear: 2-8 3,462,237  │ │ └─weight ├─3,455,488  │ │ └─bias └─6,749  │ └─Dropout: 2-9 --  ===========================================================================  Total params: 42,263,133  Trainable params: 42,263,133  Non-trainable params: 0  =========================================================================== |